مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

الموارد اللُّغويَّة الحاسوبيَّة

مباحث لغوية 01

تحرير

د. مُحسن رَشوان   د. المُعتزّ بالله السَّعيد

الباحثون:

د. عبد العاطي هوَّاري   د. المُعتزّ بالله السَّعيد

د. سامِح الأنصاري   د. مُحسن رَشوان

مباحث لغوية ٥٦

الموارد اللُّغويَّة الحاسوبيَّة

تحرير

د. مُحسن رَشوان | د. المُعتزّ بالله السَّعيد

الباحثون:

د. عبد العاطي هوَّاري | د. المُعتزّ بالله السَّعيد

د. سامِح الأنصاري | د. مُحسن رَشوان

١٤٤١هـ - ٢٠١٩م

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية

King Abdullah Bin Abdulaziz Int'l Center for The Arabic Language

مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

الموارد اللُّغويَّة الحاسوبيَّة

الطبعة الأولى
١٤٤١ هـ - ٢٠١٩ م

المملكة العربية السعودية - الرياض
ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، ١٤٤١هـ.
فهرسة مكتبة الملك فهد الوطنية أثناء النشر
رشوان، محسن
الموارد اللغوية الحاسوبية./ محسن رشوان؛ المعتز بالله السعيد .-الرياض، ١٤٤٠هـ
..ص؛ .. سم
ردمك: ٩-٥٤-٨٢٢١-٦٠٣-٩٧٨
١ - اللغة العربية - معالجة البيانات أ. السعيد ، المعتز بالله (مؤلف مشارك) ب. العنوان
ديوي ٤١٠,٢٨٥ ١٠١٦٩/ ١٤٤٠
رقم الإيداع: ١٠١٦٩/ ١٤٤٠
ردمك: ٩-٥٤-٨٢٢١-٦٠٣-٩٧٨

التصميم والإخراج
دار وجوه للنشر والتوزيع
Wojooh Publishing & Distribution House
www.wojoooh.com
المملكة العربية السعودية - الرياض
الهاتف:4562410 الفاكس:4561675
للتواصل والنشر:
info@wojoooh.com

بسم الله الرحمن الرحيم


مكتبة لسان العرب
أ. علاء الدين شوقي
@lisanarb
www.lisanarb.com

فهرس الكتاب

كلمة المركز

يعمل المركز في مجال البحث العلمي ونشر الكتب مستهدفاً التركيز على المجالات البحثية التي ما زالت بحاجة إلى تسليط الضوء عليها، وتكثيف البحث فيها، ولفت أنظار الباحثين والجهات الأكاديمية إلى أهمية استثمارها بمختلف وجوه الاستثمار، وذلك مثل مجال (التخطيط اللغوي) و (العربية في العالم) و(الأدلة والمعلومات) و (تعليم العربية لأبنائها أو لغير الناطقين بها) إلى غير ذلك من المجالات، وإن من أهم مجالات البحث المستقبلية في اللغة العربية مجال (العربية والحوسبة ، والذكاء الاصطناعي) حيث إن حياة اللغات ومستقبلها مرهونة بمدى تجاوبها مع التطورات التقنية والعالم الافتراضي، وكثافة المحتوى الالكتروني المكتوب، وهو ما يشكّل تحديا حقيقيا أمام اللغات غير المنتجة للمعرفة أو للتقنية.

وقد عمل المركز على تسليط الضوء على هذا المجال التخصصي؛ مستعينا بالكفاءات القادرة من المهتمين بالتخصص البيني (بين اللغة والحاسوب) مقدّرا جهودهم، وهادفاً إلى نشرها، وتعميم مبادئها، راغباً أن يكون هذا المسار العلمي مقررا في الجامعات في كلية العربية والحاسوب، ومجالا بحثيا يقصده الباحثون الأكديميون، والجهات البحثية العربية.

وقد أصدر المركز سابقا ستة عشر كتاباً مختصا في (حوسبة العربية) وفي الإفادة من (المدونات اللغوية) في الأبحاث العربية، ويحتفل بإصدار سبعة كتب جديدة مختصة في (حوسبة العربية والذكاء الاصطناعي)، ويقدمها للقارئ العربي، وللجهات الأكاديمية؛ للإفادة منها في مناهج التعليم والبناء عليه، وهذه الكتب السبعة هي: (العربيَّة والذَّكاء الاصطناعيّ، تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية، خوارزميات الذكاء الاصطناعي في تحليل النص العربي، مقدمة في حوسبة اللغة العربية، الموارد اللغوية الحاسوبية، المعالجة الآلية للنصوص العربية، تطبيقات أساسية في المعالجة الآلية للغة العربية).

ويشكر المركز السادة مؤلفي الكتب، ومحرريها، لما تفضلوا به من عمل علمي رصين، وأدعو الباحثين والمؤلفين إلى التواصل مع المركز لاستكمال المسيرة، وتفتيق فضاءات المعرفة.

وفق الله الجهود وسدد الرؤى.

الأمين العام

أ. د. محمود إسماعيل صالح

مقدمة

تُعَدُّ الموارِدُ اللُّغويَّةُ ركيزةً أساسيَّة لبناء وتطوير أدوات الـمُعالَجة الاليَّة للُّغات الطَّبيعيَّة؛ حيثُ تُمثِّلُ ضابطًا معياريًّا يُمكنُ الاسترشادُ به في وصف واقع اللُّغة بمُستوياتِها الـمُتعدِّدة عبرَ الزَّمان والمكان؛ وهي أيضًا وسيلةٌ لتقويم أدوات المُعالَجة الآليَّة للُّغات. أَضِف إلى ذلكَ أنَّ توظيفَ الموارد اللُّغويَّةَ في الصِّناعة الـمُعجميَّة وتطوير أدوات تعليم اللُّغة قد ساعدَ بصورةٍ كبيرةٍ في الـمُوائمة بينَ اللُّغة الموصوفة ومُستعمليها؛ كما مكَّنَ هؤلاء الـمُستعملينَ من الوُقوف على إشكالات اللُّغات الطَّبيعيَّة والتَّفكير في الوسائل النَّاجعة لـمُعالجَتِها.

إنَّنا نُقدِّمُ للقارئ العربيّ كتاب (الموارد اللُّغويَّة الحاسوبيَّة) ضِمنَ سلسلةٍ من الكُتُب الَّتي تُعنى بحوسَبة اللُّغة، آملينَ أن تُسهمَ هذه السِّلسلةُ في إثراء المكتبة العربيَّة بمصادرَ داعمةٍ ومُوجِّهةٍ للمعنيِّينَ بمُعالجَة اللُّغات الطَّبيعيَّة في ميادين البحث والصِّناعة والتَّدريس، وأن تكونَ هذه السِّلسلة باكُورةً لسلاسلَ أخرى في ذلك الحقل العلميّ الرَّحْب.

ولما كانَ الهدفُ من هذا الكتاب توجيهَ القارئ العربيّ إلى الموارد اللُّغويَّة الحاسُوبيَّة الَّتي تُمكِّنُهُ من استيعاب منطق الآلة في التَّعاطي مع البيانات اللُّغويَّة كبيرة الحجم نسبيًّا، والوُقوف على الموارد اللَّازمة لبناء وتطوير أدوات الـمُعالجة الآليَّة للُّغة الطَّبيعيَّة، فقد جاءَ في خمسة فُصُولٍ، على النَّحو الآتي:

- الفصل الأوَّل: الموارِد المعجميَّة العربيَّة الحاسُوبِيَّة؛ يُعنى بالتَّعريف بهذه الموارد وتصنيفها من حيثُ الشَّكل وطبيعة الـمُحتوى. ويُقدِّمُ كذلكَ رُؤيةً منهجيَّةً لآليَّات الإفادة من الموارد الـمُعجميَّة الحاسوبيَّة، لا سيَّما الموارد العربيَّة، في مُعالجة اللُّغات الطَّبيعيَّة وتطوير صناعة الـمُعجم العربيّ.
- الفصل الثاني: المدَوَّنات اللُّغَوِيَّة؛ ويَعرِضُ هذا الفصلُ لمفهوم الـمُدوَّنات اللُّغويَّة وأساليب بنائِها ومُعالجَتِها آليًّا، ومجالات الإفادة منها في ميادينَ عديدة، تشملُ البحثَ اللُّغويَّ ومُعالجة اللُّغات الطَّبيعيَّة والصِّناعةَ المُعجميَّة، وميادينَ أخرى.
- الفصل الثَّالث: الشَّبكات الدِّلاليَّة؛ يُقدِّمُ هذا الفصل لمحةً عن التَّحليل الدِّلاليّ للجُملة، ثُمَّ عرضًا للشَّبكات الدِّلاليَّة الَّتي تُعَدُّ موردًا أساسيًّا للتَّحليل الدِّلاليّ العميق في اللُّغات الطَّبيعيَّة. ويُعنى هذا الفصلُ بالوقوف على آليَّات توظيف الشَّبكات الدِّلاليَّة في الـمُعالجَة الآليَّة للدِّلالة.
- الفصل الرَّابع: موارد التَّعلُّم الآليّ؛ ولهذا الفصل بُعدٌ تطبيقيٌّ؛ حيثُ يعرِضُ لخمسة مواردَ أساسيَّة للتَّعلُّم الآليّ، هي: شجرة القرار، ومُصَنِّف بايز الـمُبَسَّط، والشَّبكات العصبيَّة، وآليَّات الـمُتَّجهات الدَّاعمة، ونماذج ماركوف الـمُخَبَّأة. ويعرِضُ الفصلُ لجوانب الإفادة من هذه الموارد في توجيه الآلة إلى مُحاكاة ذكاء الإنسان في فهم اللُّغة ومُعالجَة بياناتِها.
- الفصل الخامس: نمذجة اللُّغة؛ ولهذا الفصل بُعدٌ تطبيقيٌّ أيضًا؛ حيثُ يُعنى بتوظيف الموارد اللُّغويَّة في بناء النَّماذج اللُّغويَّة، عبرَ أساليبَ إحصائيَّة، مثل: النَّحو العدديّ، والتَّنعيم. ويُقدِّمُ الفصلُ أمثلةً على مجالات الإفادة من بعض هذه الأساليب في مُعالجَة اللُّغات الطَّبيعيَّة.

إنَّنا نأمل أن يجدَ القارئُ الكريمُ تنوُّعًا وثراءً بينَ جَنَبات هذا الكتاب وفي ثنايا فُصُوله؛ حيثُ تشارَكَت في تأليفِهِ نخبةٌ من المؤلِّفين الَّذينَ يجمعُونَ بينَ الخبرة الأكاديميَّة والخبرة العمليَّة في ميادين صناعة الموارد اللُّغويَّة الحاسُوبيَّة.

نسألُ الله تعالى أن يتقبَّلَ هذا الجهدَ بالذِّكر الحَسَنِ والأجرِ الجزيل، وأن يجعلَه من العلمِ الَّذي ينفعُ أصحابَه بعد مماتهم.

الـمُحرِّران

الفصل الأول

المواِرد المعجَمِيَّة العَرَبِيَّة الحاسُوبيَّة

د. عبد العاطي هوَّاري

١ – مدخل إلى المواِرد المعجَمِيَّة العَرَبِيَّة الحاسُوبيَّة.

٢ – في التعريف بالمواِرد المعجَمِيَّة الحاسوبية.

٣ – المواِرد المعجَمِيَّة ومعالجة اللغات الطبيعية.

٤ – الصناعة المعجمية الحاسوبية.

٥ – المواِرد المعجَمِيَّة العربية الحاسوبية.

٦ – الأفكار البحثية المقترحة في إطار العمل المعجمي الحاسوبي العربي.

١- مدخل إلى الموارِد المعجَمِيَّة العَرَبِيَّة الحاسُوبيَّة

يتم التمييز، في إطار الدرس المعجمي (ابن مراد، ١٩٩٨) ، (الفهري، ١٩٩٩)، (الفهري، ١٩٩٧)، بين ثلاثة مجالات تخص ثلاثة مستويات بحثية معجمية متمايزة؛ الأول: المعجم الذهني (Mental Lexicon)، ويختص بدراسة الجانب الذهني من المعجم؛ كاكتساب الثروة اللفظية وتعرُّفها وطريقة تنظيمها في الذهن، وآليات توليدها واستعمالها. والثاني هو المعجم اللغوي (Lexicon) ويقصد به مجموع الثروة اللفظية؛ الكلمات والتعابير الاصطلاحية (Idioms) الموجودة لدي مجموع المتحدثين بلغة ما. والثالث هو المعجم المصنوع أو المدوَّن (Dictionary)، ويكون محاولة لتمثيل المعجم اللغوي للغة ما في صورة مورِد معجمي وهو بذلك عمل ينتمي إلى الصناعة المعجمية (Lexicography)[1].

ويندرج موضوع هذا الفصل تحت إطار المستوى الثالث فيتناول الموارِد المعجَمِيَّة الحاسوبية؛ مفهومها، وطبيعتها، وأشكالها، وعلاقتها بمعالجة اللغات الطبيعية. بالتركيز على الموارِد المعجَمِيَّة العربية الحاسوبية، واقعها الراهن، واقتراح تصور لآفاق العمل المعجمي العربي الحاسوبي صناعةً وبحثا.

٢- في التعريف بالموارِد المعجَمِيَّة الحاسوبية

أحدثَ دخولُ الحاسوبِ مجالَ العمل المعجميّ -صناعةً وبحثًا- ثورةً شاملة في تقنيات الصناعة المعجمية، أعقبتها ثورة مماثلة في المفاهيم والمعتقدات والتقاليد المعجمية. فحدثت تبدُّلات كبيرة في أولويات العمل في هذا المجال تخطيطًا وتنفيذًا وتحديثًا. علاوة على اختلاف غير قليل في الأهداف الصناعية والبحثية. ويمكن أن نتصور آثار هذه الثورة في مستويات متعددة[2]؛ في تقنيات العمل المعجمي التقليدي: منهجياته وإجراءاته وأدواته وبنائه. وفي إيجاد أشكالٍ لمعاجم أو موارِد مُعجَمِيَّة جديدة. وفي طرائق التعامل معها واستعمالها وفي توظيفها. وأيضًا في ظهور مجالات بحثية

١- على أن هناك من يُسَوِّى بين المصطلحين (Lexicon و Dictionary) في العمل المعجمي الغربي أيضا، فيستخدمهما مترادفين.

٢- علاوة على بَدَهيات جدوى استعمال الحاسوب في معالجة أي مادة في أي مجال معرفي؛ من سرعة وإنجازيه عالية ودقة وسعة تخزينية فائقة وإمكانية متابعة التحديث، إضافة إلى طريقة تقديم المادة واستدعائها.

معجمية جديدة، واقتراح إجراءات جديدة للبحث المعجميّ ومنطلقاته وغاياته[١]. ويذهب بعض الباحثين إلى أن ثورة الحاسوب لم تغير في الطريقة التي يُجرَى بها البحث فحسب، بل إنها أيضًا فتحت آفاقًا لحقول بحثية جديدة تهدف إلى فهم للعقل البشرى على ضخامتــه وتعقيد.

وينبغي التمييز بين نمطين من الموارِد المعجمية الحاسوبية: الأول: موارِد معجمية للمستعمل البشري تكون تطبيقا حاسوبيًّا قائما بذاته، مثل المعاجم الإلكترونية، تعرض مادتها في صورة واجهة على شاشة الحاسب، تُسَهِّل عمليةَ البحث عن الكلمة ومعلوماتها اللغوية؛ والآخَر: موارِد معجمية تُجعل لأنظمة الحاسوب الداعمة لمعالجة اللغات الطبيعية، فمستعملو هذه الأنظمة الحاسوبية لا يتعاملون مباشرة مع المورِد المعجمــي، بــل يتعاملون مع التطبيقات المبنية على هذه الموارِد ويكون جزءً من نظام أكبر كما في المدقق الهجائي الخاص بمعالج الكلمات أو في المعجم المصطلحي لنظم مساعدة المترجمين.

٢ , ١ – أشكال الموارِد المعجَمِيَّة الحاسوبية

تتنوع الموارِد المعجَمِيَّة الحاسوبية فتشمل أشكالا متباينة طبيعة، وحجما، وغاية، وشكلا نهائيًّا تتجلى فيه. ويمكن استعراض أشكال الموارِد المعجَمِيَّة من خلال تصنيفها من حيث شكلها الحاسوبي الذي تتجلى فيه، ومن حيث طبيعة المحتوى المعجمي الدلالي الذي تقدمه، ومن حيث غاياتها.

- **أشكال الموارِد المعجَمِيَّة الحاسوبية من حيث الشكل**

  - **المعجم المقروء آليًّا**

يُعدُّ المعجمُ المقروءُ آليًّا[٢] نسخةً حاسوبيةً من طبعته الورقية، أو تمثيلاً حاسوبيًّا للمعجم الورقي/ التقليدي يظهر في هيئة إلكترونية تسمح للآلة/ الحاسوب بالقيام

١– وقد حدا كل ذلك بجُلّ مَن تصدَّى للتأريخ للمعجم بأن يقسم عمله إلى قسمين: ما قبل الحاسوب (ويروق لكثير من الباحثين تسميتها المعجمية التقليدية)، وما بعد الحاسوب (وتسمى المعجمية الحاسوبية).

٢– يحترز البعض على عبارة «المقروء آليا» بقولهم إنه ليس المقصود أن الحاسوب يقرأ المعجم بل فقط أن المعجم في هيئة الكترونية تسمح للآلة / الحاسوب أن يقوم بمعالجات عليها (Litkowski, 2005).

بمعالجات عليها. وإلى جانب تمثيله للمعجم المطبوع فإنه يختلف عنه باحتوائه معلوماتٍ لغوية لا تظهر في المعجم المطبوع نظرا لاختلاف طبيعة الآلة عن المستخدم البشري.

وتاريخيًّا لم يرتبط نوعٌ من أنواع الموارِد المعجَمِيَّة الحاسوبية بالمعجمية الحاسوبية قدر ارتباط المعجم المقروء آليًّا بها. حتى أن مصطلحَ المعجمية الحاسوبية نفسِه قد ظهر أول ما ظهرَ على يد «أمسلر»(Amsler) من خلال دراسته عن بنية معجم ويبستر السابع، وكان يعني دارسة المعجم المقروء آليًّا. وقد بدأ المعجم المقروء آليًّا في الظهور في منتصف الستينيات وازداد الاهتمام به مع بداية التسعينات من القرن العشرين. ولعل أشهرَ نماذجِه معجم لونجمان للإنجليزية المعاصرة[1] LDOCE.

ويُنظَر للمعاجم المقروءة آليًّا باعتبارها مورِدًا قيِّمًا للمعلومات اللغوية المستخدمة في مجال معالجة اللغات الطبيعية، وذلك لاحتوائها جُلّ المعارف اللغوية والدلالية (Pustejovsky & Boguraev, 1993). فمن مادته يَسْتَخْرِج اللسانيون الحاسوبيون المعلوماتِ اللغوية دلاليةً وتركيبيةً وصرفيةً، يتم توظيفها في مجال معالجة اللغات الطبيعية، كما أن هذه المعارف اللسانية الموجودة في المعجم المقروء آليا تعد مادة ملائمة لاشتقاق قواعد للمعارف منها. فلقد كانت نتائج الأبحاث المبكرة في مجال المعجم المقروء آليا مبشرة فقادت كثيرين إلى الشعور بأن قواعد معارف ضخمة يمكن أن تشتق بسهولة اشتقاقا آليًّا من المعاجم المقروءة آليًّا (Ide & Véronis,1994).

ولقد وجد مطورو النظمِ المعجميةِ الحاسوبيةِ والمعجميون الحاسوبيون أن المعجمَ المقروءَ آليًّا لا يفي بمطالب استخلاص المعلومات بالشكل الذي يرضونه فأخذوا في ابتكار الأدوات البرمجية لتحويله إلى معاجم قابلة للتوسيع آليًّا لتصبح في صورة صالحة لمعالجة اللغة الطبيعية مباشرة، وتستخدم في ذلك أدوات إحصائية للمفردات وعلاقاتها في MRD للخروج بشبكة دلالية تحكم بنيتها، أو بتحليل كلمات التعريفات المعجمية للوصول إلى المعاني النووية التي تحكم بنية المعجم، أو بتحليل التعريفات نفسها وهي الإجراءات التي طبقها فريق عمل على معجم لونجمان المقروء آليًّا لتحويله لمعجم قابل للتوسيع آليًّا صالح لتطبيقات معالجة اللغة (Svensén, 1993).

1- http://www.pearsonlongman.com/ldoce/.

- **قاعدة البيانات المعجمية**

هي صياغة للمادة المعجمية (المداخل المعجمية والمعارف اللغوية المتعلقة بها) في صورة قاعدة بيانات، بما لقواعد البيانات من إمكانات في التخزين، والضبط، والربط العلائقي، وإمكانات البحث، والفهرسة، والاستخلاص، والإحصاء. ويَعُدُّ البعضُ قاعدة البيانات المعجمية نسخة من المعجم المقروء آليًّا، غير أنها نسخة معدلة الأخطاء تتجاوز التضاربات الداخلية التي قد تكون موجودة في المعجم المقروء آليًّا. غير أن الفروق بينهما على المستوى التنظيمي المعجمي ليس بالقليل ويبرر الحديث المستقل عن كل منهما.

ويمكن النظر إلى قواعد بيانات معجمية بوصفها مخزنا هائلا للثروة اللفظية؛ ألفاظا ومعلومات متعلقة بها، مصوغة في صورة منظومية، يمكن توظيف محتواها في بناء موارِد مُعجَمِيَّة أخرى أو برمجيات حاسوبية فيما يخص معالجة اللغات الطبيعية.

وتعد قاعدة بيانات المعجم الإيطالي من الأعمال الأكثر شهرة في سياق الحديث عن قواعد البيانات المعجمية؛ فلقد قامت منهجية تمثيل المحتوى المعجمي الدلالي لقاعدة البيانات المعجمية الإيطالية على مقولات التوجه العلائقي (Relational Approach) في تمثيل المعنى. فقد زخرت قاعدة بيانات المعجم الإيطالي بالعلاقات على أشكالها المختلفة، وهو الأمر الذي لم يكن معهودا في معاجم التعريفات، للدرجة التي جعلت بعض الباحثين يشير إلى إمكانية دمج المعجم والمكنز.

- **أشكال الموارِد المعجَمِيَّة من حيث طبيعة المحتوى المعجمي ومنهجية تمثيله**

يمكن تصنيف الموارِد المعجَمِيَّة من حيث طبيعة المحتوى المعجمي الدلالي ومنهجية تمثيله إلى قسمين: الأول للموارِد المعجمية، وهي الموارِد التي تركز على جوانب المعنى والاستعمال؛ والقسم الآخَر هو الموارِد المعجَمِيَّة الدلالية ويقصد بها الموارِد المعجَمِيَّة التي يركز محتواها على الجوانب المتعلقة بتمثيل الأبنية التركيبية للوحدات المعجمية وتنميط سلوكها التركيبي.

- **الموارد المعجمية**

تركز الموارِد المعجَمِيَّة على تمثيل المعنى من خلال إحدى منهجيتين:

- الموارِد المعجَمِيَّة التعريفية: (Dictionary) وهي التي تعتمد منهجية التعريف المعجمي (Lexical Definition) في تمثيل المعلومات والمعارف اللغوية.

ويمكن هنا أن نضرب مثالا على هذا النوع بمعجم لونجمان للإنجليزية المعاصرة (Longman Dictionary Of Contemporary English): وهو معجمٌ قد أخذ اعتناءً من قبل اللسانيين واللسانيين الحاسوبيين[1]. فالحديث عن معجم لونجمان لا يعد حديثا عن تجربة معجم بقدر ما هو مراجعة لسلسلة من الدراسات المعجمية الدلالية لفريق عمل متكامل من المتخصصين.

وقد أقام المعجم منهجيته في تحليل المحتوى المعجمي الدلالي وتمثيله على فكرة أساسية، هي استعمال قائمة كلمات محددة[2] بمعان محددة، يتم تعريف بقية مداخل المعجم بها، وقوائم من التصنيفات التركيبية والدلالية، وهو ما جعله نموذجا مميزا للتطبيقات الحاسوبية في مجال معالجة اللغات الطبيعية من خلال المعلومات التي يقدمها مثل عمله تراتبية للأسماء، والتصنيف الاستعمالي، وبيان حقول الاستعمال، وتوضيح القيود الانتقائية للوحدات المعجمية المعَرَّفَة.

ولا يخفى عمق التحليل المعجمي الدلالي الذي وقف وراء هذا العمل، وفاعلية لغة التعريف بالتحكم في مفرداتها وأبنيتها. غير أن كل هذه المزايا التي وجدها الدارسون في معجم لونجمان لم تُعْفِهِ من نقد توجه إلى منهجيته النظرية وتطبيقاته.

- موارِد مُعجَمِيَّة علائقية شبكية: وهي في فلسفة بنائها وتمثيلها للمعنى أقرب إلى المكانز اللغوية التي تعتمد في بنائها على مقاربة العلاقات الدلالية القارة في المعجم اللغوي، غير أنها توظف تقنيات العمل الحاسوبي في توثيق عُرَى العلاقات بين مفردات وعبارات المعجم.

١- أصبح من المعتاد في مجال معالجة اللغة الإنجليزية آليًّا الاعتماد على مادة معجم لونجمان في نسخته الالكترونية. ومما ينبغي إيراده هنا أن للمعجم نسختين: الأولى ورقية، والثانية حاسوبية في صورة معجم مقروء آليًّا وأن بين النسختين كثيرا من الاختلافات.

٢- تتكون هذه القائمة من ٢٣٠٠ كلمة، إضافة إلى ٣٠٠٠ مشتق من مشتقاتها.

ويمكن أن نمثل على هذا النوع من الموارِد بشبكة الكلمات WordNet [١]، وهي تقوم على أسس نفسية لسانية بالأساس وتعتمد في تنظيمها على فكــرة المكنز، وتحاول شبكة الكلمات تمثيل المـادة المعجمية تمثيلاً يشابه طريقة تنظيم العقل البشريّ للمادَّة المعجميـة «فالبنية الهرمية التي تطور عن طريق نظريات تنظيم المعرفة البشرية تقدم مادة مفيدة لمشروعات أبحاث الذكاء الاصطناعي.

ورغم كل ما في شبكة الكلمات من مزايا وطاقات تَعِدُ بتطبيقات في مجال معالجة اللغات الطبيعية؛ فإنها لا تعدم من يرى فيها نقائص مثل صعوبة الربط القائم بين أقسام الكلام المختلفة، وقلة التَّعبيرات الاصطلاحية المدرجة، ناهيك عن أن تصنيفها لم يخطط بشكل شامل ومحكم.

- **موارِد مُعجَمِيَّة دلالية (Lexical Semantic Resources)**

تنطلق الموارِد المعجَمِيَّة الدلالية من نظريات معجمية دلالية، فتطبق بعضا من فروضها النظرية على معاجم اللغة، فتقدم تمثيلا دلاليا وتركيبيا للوحدات المعجمية. وهي موارِد مُعجَمِيَّة دلالية تهدف إلى تمثيل الأبنية التركيبية والسلوك التركيبي للوحدات المعجمية إلى جانب تمثيل المحتوى الدلالي لهذه الوحدات المعجمية، خصوصا الأفعال والمشتقات وتقوم بتصنيفها تصنيفا تركيبيا ودلاليا في آن. وهذه الموارِد لها أهميتها الكبيرة في تطبيقات معالجة اللغات الطبيعية إذْ إنها أعمق تحليلا وألصق بالسياقات التركيبية المحتملة للكلمة. ومن أمثلة ذلك: شبكة الأطر (FrameNet) وشبكة الأفعال (VerbNet) وبنك الأبنية الحملية (PropBank).

وسنمثل لاثنتين منهما فيما يلي.

شبكة الأُطُر

تعد شبكة الأطُر[٢] المعجمية الدلالية من أهم الأعمال في مجال بناء الموارِد المعجَمِيَّة الدلالية على مستوى التأسيس النظري خصوصا؛ إذ قد تم بناؤها على هدى من نظرية

١- رابط المشروع للاطلاع: http://wordnet.princeton.edu.

٢- عنوان مشروع FrameNet ورابطه على شبكة الإنترنت هو: http://framenet.icsi.berkeley.edu ويغطي المشروع عدة لغات إلى جانب الإنجليزية وهي الألمانية واليابانية والأسبانية.

فيلمور (Fillmore) عن نحو الحالة (Case Grammar) والأدوار الدلالية/ المحورية؛ فالأدوار الدلالية (Semantic Roles) تلعب دورا مهما في هذا العمل. وتعد شبكة الأطر مورِدًا معجميا دلاليا غايته بالأساس تنظيم المعارف المعجمية على أسس تركيبية ودلالية لتحقيق أغراض المعالجة الآلية للغات الطبيعية، إضافة إلى الجوانب النظرية المتمثلة في التحليلات المعمقة للبنية المعجمية الدلالية للمعجم.

والإطار الدلالي عبارة عن بنية عامة تتكون من مجموعة من العناصر؛ تبدأ بتعريف عبارة عن توصيف للمفهوم أو المعنى، ثم مجموعة من العناصر (ويمكن النظر إليها باعتبارها تحقيقًا لفكرة الأدوار الدلالية كما في نظرية نحو الحالة لفيلمور) هذه العناصر منها ما هو أساسي ومنها ما هو غير أساسي. ويكون دور العَنونة الدلالية هو ربط كل قراءة دلالية لوحدة معجمية بالإطار المناسب له من الأطر المحددة سلفا، إضافة إلى تحديد العناصر الأساسية والعناصر غير الأساسية لكل كلمة. ويتأسس المشروع على ما يقرب من ٨٠٠ إطار معجمي دلالي يتم تصنيف الوحدات المعجمية من خلالها.

بنك الأبنية الحملية (PropBank)

يتوجه مشروع بروب-بانك إلى تحليل البنية الحملية (Argument Structure) للفعل والمشتقات في عدد من اللغات (الإنجليزية والعربية والصينية والهندية. (وغايته توصيف السلوك التركيبي مربوطا بالدلالة/ المعنى لكل فعل أو مشتق، ومن جهة أخرى إنجاز تصنيف معجمي دلالي وتركيبي لمعجم كل لغة.

ويتم التحليل (Palmer, 2005) بالبدء بتحليل الدلالات الممكنة لكل فعل أو مشتق من خلال أمثلة محللة تركيبيا تحليلا شجريا وتقديم تعريف مبسط له وتحديد البنية الحملية الممكنة مع كل دلالة وتوضيح المكملات الجملية الممكنة مع ربط الجمل/ العبارات الموجودة في المدونة بالتحليل، فتصبح المخرجات النهائية عبارة عن تحليل معجمي دلالي لأفعال ومشتقات المدونة يقدم توصيفا للسلوك التركيبي للفعل أو مشتقاته وفصل متعدد المعنى على أسس تركيبية ودلالية.

٢,٢- روافد تطوير العمل المعجمي الحاسوبي

تسترفد الموارِد المعجَمِيَّة الحاسوبية رافدين يتشاركان في تطويره؛ الأول رافد لساني والثاني رافد حاسوبي. ويهمنا في هذا السياق أن نتناول الرافد اللساني للتطوير. فالعمل المعجمي الحاسوبي يرتبط في عمق تحليله بالتأسيس النظري اللساني وهو ما يمكن ملاحظته، على سبيل المثال، في تأثير أعمال بيث ليفين Levin المتعلقة بالتصنيف المعجمي الدلالي للأفعال (Levin ١٩٩٣) في بناء الموارِد المعجَمِيَّة الحاسوبية وفي تنظيم محتواها الدلالي والتركيبي.

وفي هذا السياق نعرض لنموذجين فيما يتعلق بالتنظير اللساني للتطوير في مجال بناء الموارِد المعجَمِيَّة الحاسوبية؛ هما نموذج معجم ميلتشوك، والثاني نموذج نظرية المعجم التوليدي لبوسطيوفسكي[1].

- **نموذج معجم ميلتشوك (Explanatory Combinatorial Dictionary)**

يمكن النظر إلى تصور ميلتشوك للمعجم باعتباره أحد أنضج المنهجيات التي قدمت في العصر الحديث فيما يخص تمثيل البنية الدلالية لمعجم؛ وذلك لارتكازها على تصورات لسانية نظرية معمقة، دون تجاهل لما تقدمه نتائج تحليل المدونة النصية.

وتركز منهجية معجم ميلتشوك على جوانب المحتوى: فهو معجم لتمثيل المعنى وتوضيح آليات تآلف الوحدات المعجمية. ويقوم تمثيل معنى الكلمة في المعجم على محورين يسميهما الوظائف المعجمية؛ هما العلاقات الرأسية (Paradigmatic Functions) (أي العلاقات الدلالية بين الوحدات المعجمية) والعلاقات الأفقية (Syntagmatic Functions) (تآلف الكلمة مع جاراتها في العبارة أو الجملة) وذلك بغرض تتميم التوصيف الدلالي للكلمة المدخل.

وتتكون بنية تعريف الوحدة المعجمية في تصور ميلتشوك من مجموعة من القوالب الفرعية (Mel’ˇcuk, 1988 & 1995)، هي: المكونات النموذجية للتعريف. وهي عبارة عن قالب ثابت يتم الالتزام به في أي تعريف. والمكونات العامة، وهي التي تحدد

١- ينبغي الإشارة إلى أن نموذج ميلتشوك قد صار بالفعل تطبيقيا واقعيا متحققا، غير أن نظرية بوسطيوفسكي ظلت - حتى الآن - قيد التحقيق، وإنْ تم الإفادة منها في مجالات مختلفة في تحليل اللغة وتطبيقات معالجة اللغات الطبيعية.

الفئة التركيبية التي تنتمي إليها الوحدة المعجمي. والمكونات الضعيفة، وهي أجزاء من التعريف يمكن الاستغناء عنها في بعض السياقات الدلالية. والمكونات الاختيارية للتعريف وهي التي يمكن تحييدها في التعريف، وإنما يؤتى بها فقط لتسييق الكلمة. والقيود، وهي السمات التي تميز بين تعريفات الكلمات. والمكونات الجاهزة، ويقصد بها السمات المصوغة قبلا بحيث تضاف للتعريف بوصفها وحدات تعريفية سابقة التجهيز.

وتأتي أهمية هذا العمل من كونه تطبيقا لنظرية لسانية دلالية من جهة، ومن كونه يتوجه إلى تطبيقات معالجة اللغة آليًّا من جهة أخرى. كما أنه قد قدم تأسيسًا نظريًّا لبنية المعجم يقوم على أسس دلالية، إضافة إلى ما قدمه في مجال الوحدات المعجمية متعددة الكلمات.

- **نموذج المعجم التوليدي لبوسطيوفسكي (Generative Lexicon)**

تصنَّف نظرية بوسطيوفسكي بوصفها نظرية في الدلالة المعجمية الحاسوبية للكلمة، وهي محاولة لاقت قبولا واستحسانا في الأوساط اللسانية الحاسوبية باعتبارها نظرية لتمثيل المعرفة المعجمية، تُحاول تجاوزَ المنجز في التصورات السابقة، مفيدة من النظريات السابقة.

ينطلق بوسطيوفسكي (Pustejovsky,1995) من نقد سكونية (استاتيكيَّة) المقترحات السابقة في معالجة المعجم، فقد رأى أن الأفكار التي بنيت عليها المعاجم أفكار سكونية في تمثيلها المحتوى المعجمي الدلالي، يتم سرد الدلالات فيها دون روابط قوية ودون التركيز على الطاقة التوليدية الكامنة في الكلمات. فكان التفكير في أن يجعل المعالجة تفاعلية تمكن من التعامل مع الكلمات في سياقات جديدة أو مختلفة عما هو وارد في المعجم، وفي المقابل يقترح تصورا حركيا تفاعليا يمكن أن يسهم في مجابهة الاستعمال الإبداعي/ المتجدد للكلمة في نصوص جديدة وفك الالتباس. فالكلمة يتحدد معناها حسب السياق، والسياقات التي (قد) ترد فيها الكلمة سياقات لا نهائية، فلا يمكن محاصرتها مهما أمكننا السيطرة على المدونة من حيث ضخامة الحجم، والتحديث الدائم، ولكن يمكن السيطرة على آلية توليد دلالة الكلمة في السياقات / أنماط السياقات.

ويفترض بوسطيوفسكي، في سبيل تمثيل الطاقة التوليدية للكلمة، أربعة أبنية لتمثيل الوحدة المعجمية:

- البنية الحملية (Argument Structure) وهي تتعلق بالتحقق التركيبي للكلمة (عدد المحمولات وأنماطها التي تتحقق معها في المستوى التركيبي).
- البنية الحدثية (Event structure) تعين نمط الحدث في الفعل، وهو يعد بمثابة تصنيف لطبيعة الفعل أو المشتقات الفعلية لتقديم خصائصها الجهية.
- بنية السمات (الكواليا) (Qualia structure) تقدم السمات الأساسية للوحدة المعجمية التي تتحكم في آليا تراكبها مع الوحدات المعجمية الأخرى.
- بنية التوارث (Lexical inheritance) وتتعلق بالبنية الكلية لمعجم لغة، فتحدد الطرائق التي تترابط بها الكلمات دلاليا واشتقاقيا فيما بينها في داخل المعجم.

ويمكن النظر إلى تصور بوسطيوفسكي باعتبارها التصور الأوجه، حاليا، في مجال تمثيل المعارف المعجمية الدلالية بغرض الاستخدام في مجال معالجة اللغات الطبيعية لكونه تصور ديناميكي للمعجم يتصدى للمعاني أو الاستعمالات الجديدة للكلمة، فلقد أصبح التركيز موجها إلى الطاقة التوليدية للوحدة المعجمية وكيفية تآلفها أو تراكبها مع وحدات معجمية أخرى، لا إلى مجرد سرد المعاني المختلفة للكلمة، والتفسير التوليدي للتآلف الممكن بين المركبات الاسمية.

٣- الموارِد المعجَمِيَّة ومعالجة اللغات الطبيعية

ترجع أهمية الموارِد المعجَمِيَّة لمجال العمل في معالجة اللغات الطبيعية، أبحاثا وتطبيقات، إلى طبيعة المعلومات التي يشتمل عليها المورِد المعجمي. والموارِد المعجمية رغم احتوائها أنماطا مختلفة من المعلومات التوصيفية لمفردات وتعبيرات؛ لا يزال قاصرا عن تلبية متطلبات معالجات اللغة الطبيعية، لذا فإننا نجد كثيرا من الأدبيات التي تصدت للحديث عن المعجم في علاقته بمعالجة اللغات الطبيعية تتصدرها جملة مثل إن المعجم ليعد عنق الزجاجة بالنسبة لمعالجة اللغات الطبيعية والذكاء الاصطناعي (Zernik, 1991).

٣,١- أنواع المعارف اللغوية التي تتطَلَّبُها معالجة اللغات الطبيعية

يقر جُلُّ اللسانيين الحاسوبيين بمركزية الموارِد المعجَمِيَّة في العمل اللساني الحاسوبي، بحثا وتطبيقات، ويرون أن المعاجم المقروءة آليًّا قد غدت موردًا ملائما للمعلومات المستخدمة في معالجة اللغات الطبيعية لاحتوائها على كمية ضخمة من المعارف المعجمية والدلالية المجموعة عبر سنوات من الجهد المعجمي (Ide & Véronis,1994).

وتتنوع المعارف اللغوية التي تطلبها أنظمة معالجة اللغات الطبيعية في الطبيعة وفي العمق[1] وهذه المعارف يقوم المعجميون بجمعها وتحليلها، وتمثيلها في سرد معجمي، وإدراجها في بِنَى مداخل المعجم بطرائق منظومية منضبطة يمكن للحاسوب أن يتعرفَ عليها، إضافة إلى تصنيف الوحدات المعجمية علي أساسها لتوظيفها في أنظمة معالجة اللغة الطبيعية.

٣,٢- متطلبات معالجة اللغات الطبيعية في الموارِد المعجَمِيَّة

ثَمَّةَ مُتَطَلَّبات لأنظمة معالجة اللغات الطبيعية، ينبغي أن تتوافر كلها أو بعضها في الموارِد المعجَمِيَّة حتى يمكن الإفادة منها آليًّا. ويمكن تصنيف هذه المتطلبات في مجموعتين؛ الأولى هي متطلبات لسانية وتتعلق بطبيعة المعارف اللسانية المقدمة وعمق تحليلها ودقة لغة تمثيلها. والثانية هي متطلبات حاسوبية وتتعلق بالشكل الذي تعرض عليه المعارف اللسانية بما يسهل إجراءات إمكانية استخلاصها واكتسابها آليًّا.

- **المتطلبات اللسانية**

كلما كان المحتوى اللساني معمقا ودقيقا ساعد ذلك في تحليلٍ معجميٍ أمثلَ للغة، وتطبيقات معجمية حاسوبية أكثرَ دقةً وفاعليةً. ويحكم درجة عمق هذا المحتوى اللساني مقدار التحليل المعجمي الدلالي للبنية الدلالية للمعجم، ودقة المنهج وقدرته على تمثيل اللغة. والمتطلبات اللسانية التي يرنو إليها العمل اللساني الحاسوبي هي تلك المعلومات التي تمكن من إنجاز المهام الآتية آليا:

١- يحدد بوجريف Bran Boguraev وتيد بريسكو Ted Briscoe المعارف اللازمة لأنظمة معالجة اللغات الطبيعية بأنها: المعارف النطقية (الفونولوجية) والصرفية والتركيبية والدلالية والبرجماتيَّة.

- **تحليل اللغة واكتسابها آليًّا**: بما تتيحه البرامج الحاسوبية من أدوات تحلل الصوت والحرف والكلمة والجملة والنص. ولا يخفى هنا أهمية المـوارِد المعجَمِيَّة في إنجاز تحليل الجوانب المرتبطة بالدلالة ابتداء بالكلمة ثم الجملة فالنص.

- **توليد اللغة**: ويقصد به تمكين البرنامج الحاسوبي من توليد اللغة؛ كلمات وجملا وعبارات تحمل دلالة.

- **فهم اللغة**: وهو موضوع أدْخِل إلى اللسانيات الحاسوبية عن طريق ازدياد الاهتمام به في مجال الذكاء الاصطناعي.

- ويمكن هنا أن نمثل على توظيف الموارِد المعجَمِيَّة في التطبيقات الحاسوبية بقاعدة المعارف المعجمية. إذ تُعَدُّ إحدى التطبيقات التي توظف الموارِد المعجَمِيَّة وأكثرها تطورًا وأعقدها بنية فمتطلبات بنائها المعجمية والحاسوبية- خصوصا المعجمية - أكبر من تلك المتطلبات التي يحتاجها أي من الموارِد المعجَمِيَّة الأخرى البسيطة[1]. كما تعد موردًا مهمًّا من موارِد المعلوماتية الحديثة وموارِد معالجة اللغات. فيحتاج بناءُ قاعدة معارف إلى كمية كلمات وتعبيرات ضخمة منظمة ومصنفة تصنيفا مُحكما، يقوم على تجلية أبنيتها الدلالية والتركيبية، هذا التصنيف يوجد في الموارِد المعجَمِيَّة الأخرى صريحا أو ضمنيًّا في بنية التعريف المعجمي، كما «تحتاج إلى بيان لنسب تردد الاستعمالات الخاصة بكل كلمة، وبيان الروابط الجلية بين الدلالات: معاني ومفاهيم ومترادفات وفروقا لغوية، وتبيان الروابط بين أقسام الكلام والتعريفات، والتصنيفات الفرعية (Jarmasz & Szpakowicz, 2001). وهناك أنواع من المعلومات يجب أن تتضمنها قاعدة المعارف وقد لا تضمن في المعجم المقروء آليًّا، وهي الجوانب السياقية للوحدات المعجمية مقدمة بشكل موسع، والمعارف الموسوعية (غير اللغوية) (Véronis, 1991).

ولا شك أن طبيعة المعلومات ومنهجية تقديمها لهما الأهمية الكبرى في هذا السياق، لذا يعد تطوير منهجيات بناء الموارِد المعجَمِيَّة تطويرًا لمجال العمل اللساني الحاسوبي كله.

---

١ - يعد البعض قواعد المعارف المعجمية أحد أشكال الموارِد المعجَمِيَّة.

- **المتطلبات الحاسوبية**

أما المتطلبات المتعلقة بالشكل (اللغوي الحاسوبي) فتتوقف عليها درجة صلاحية النصوص المعجمية للمعالجة الحاسوبية تعرُّفا واستخلاصا. فكلما كان الشكل المقدم من خلاله هذه المعلومات مطردا ومنتظما سهل ذلك في التناول الحاسوبي لهذه المادة وتوظيفها في العمل اللساني الحاسوبي.

ومن هذه المتطلبات ما هو حاسوبي كالشكل الذي تخزن فيه الموارِد المعجَمِيَّة نفسها كأن يتخذ المورِد المعجمي شكل المعجم المقروء آليًّا، أو قاعدة بيانات معجمية أو غير ذلك.

إضافة إلى طريقة تمثيلها في صورة نص معجمي تقليدي أو شبكات دلالية.

على أنه يمكن عرض جانب من السمات التي ينبغي أن تتوافر في المورِد المعجَميّ كي يكون أكثر صلاحية وفاعلية لدى المعالجة الحاسوبية، ويمكن اعتداد هذه السمات من معايير الحكم على درجة صلاحية مورِد مُعجَمِيّ للمعالجة الحاسوبية:

- **التحليلية**: أي قابلية مادة المورِد المعجمي للخضوع للتحليل سواء التحليل على مُستوى الكلمة أم التحليل على مُستوى الجُملة، وهو ما يعني وجود قواعد تركيبية مسبقة للتحرير تحكم عمل محرريه.

- **التوليدية أو التركيبية (القابلية للتوليد)**: بأن يبنى التمثيل المعجمي في المورِد المعجمي من وحدات ذرية (Atomic units) تمثل الوحدات الصغرى في البنية المعجمية الدلالية، بحيث تمكن من تعرف وتوليد الجوانب المطردة داخل البنية الدلالية المعجم. ومثال ذلك معجم لونجمان الذي اقتصرت تعريفاته على استخدام ألفَيْ كلمة تم اختيارها على أساس الشهرة والبساطة. وهو المعيار الذي تفضل معظم التطبيقات في مجال اللسانيات الحاسوبية وأنظمة معالجة اللغة الإنجليزية آليا؛ أن تعتمد عليه وتتخذه مورِدَها المعجمي. وذلك لعدة أسباب أهمها التزامه بقائمة كلمات لا يتجاوزها في التعريف. وهو أمر ذو جدوى كبيرة في تطبيقات معالجة اللغات الطبيعية، لأن قائمة الكلمات التعريفية تجعل من الممكن التعامل حاسوبيا مع النصوص التعريفية من خلال التحليل، والتوليد، والفهم: فالفكر المعتمد على تذرية المكونات أكثر اتساقا والفكر الحاسوبي.

- **النظامية (Systematic)**: وتعني هنا عرض المادة في اتساق، لا تضارب بين مكوناتها الجزئية «فالموارِد المعجمية مصدر ثري للمعلومات الدلالية المفصلة، غير أن معلوماتها الدلالية يجب أن تُنَظَّم تنظيمًا نسقيًّا (Litkowski, 2005). وتتجلى النظامية أو النسقية أيضا في الصياغة البنيوية للمعجم: وتعني الصياغة البنيوية للمعجم صياغةً تجلي الاطرادات المعجمية الدلالية، وتُظْهِر العلاقات بين الكيانات الدلالية في المعجم. وتعد (النظامية أو النسقية) أهم متطلبات المعالجة الحاسوبية لمحتوى المعجم.

٣,٣- توظيف الموارِد المعجَمِيَّة في مجال معالجة اللغات الطبيعية

وتظهر أهمية الموارِد المعجَمِيَّة لدى الحديث عما ينتج من تطبيقات معجمية في مجال تطبيقات معالجة اللغة الطبيعية، مثل البرامج المكتبية (كمعالج الكلمات) وتحليل الكلام، وتركيبه، والتلخيص الآلي، والفهرسة، واستخلاص المعلومات، والترجمة الآلية، والتحليل التركيبي الجُمَلي وغيرها، إضافة إلى العون الذي يمكن أن تقدمه المعجمية الحاسوبية في مجال تعليم اللغات (في اختيار المادة المقدمة، وطرائق تقديمها بتقنيات تفاعلية)، وفي مجال العمل المصطلحي، كإنشاء بنوك المصطلحات، والمولدات الآلية للمصطلحات، والأنظمة الخبيرة. ولقد تعددت التطبيقات للمعجمية الحاسوبية حتى غزت – أيضا – الفلسفاتِ والعلومَ والمعارفَ المختلفة.

٤- الصناعة المعجمية الحاسوبية

ينبغي هنا أن نميز بين مفهومين في إطار العلاقة بين المعجم والتقنيات الحاسوبية:

الأول: هو المعجمية الحاسوبية؛ ويقصد بها صناعة المعاجم باستخدام تقنيات الحاسوب وقدراته في التخزين والتحليل والاستفسار، ابتداء من الاعتماد على المدونات المحوسبة والأشكال الحاسوبية للتخزين مثل قواعد البيانات.

الثاني: هو المعجم الحاسوبي؛ ويقصد به المعجم المبني على أسس مفاهيمية حاسوبية تتعدى مجرد استخدام الأدوات الحاسوبية في التحليل أو التخزين أو تيسير الاستدعاء. فنقل معاجم تقليدية وتخزينها حاسوبيا لا ينتج معجما حاسوبيًّا حقيقيًّا، ولكنه ينتج معجمًا تقليديًّا مُرتديًا الثوبَ الحاسوبيّ.

ويمكن تقديم لمحة عن المساهمات الحاسوبية في بناء معجم فيما يلي:

يوظف المعجميون الحاسوبيون البرامج والأدوات الحاسوبية في إجراء عمليات الجمع والاكتساب والإحصاء والفهرسة والتحليل والتصنيف. ثم يأتي دور المعجميّ الذي يقوم بعمليات تحليل النصوص أو السياقات التي وردت فيها الكلمة أو التعبير للتوصل إلى الدلالات، ثم تصنيفها إلى دلالات مركزية ودلالات هامشية ودلالات مجازية، ثم يقوم بالعَنونة.

تنظيم المادة وتخزينها حاسوبيا: فتحول المادة المعالجة إلى صورة قاعدة بيانات معجمية تمهيدًا لاستخلاص مادة المورِد/ الموارد المعجمية منها، وللحاسوب الدور الأعظم في تنسيق وتنظيم المادة. ويذكر فريق عمل معجم «كوبيلد» (COBUILD) للحاسوب، في هذا الجانب من تقنيات العمل، أنه قد قام بفرز الكلمات بطرائق متنوعة لتصل المعلومات الخاصة بكل كلمة إلى فريق من المحررين والمؤلفين الذين يقومون بدورهم بدراسة هذه الكلمات لإنشاء ملف مفصل لمعانيها واستخداماتها في قاعدة بيانات معجمية لتصبح بالتالي المصدر الأولى لعائلة من الكتب. إضافة إلى وجوب ربط معلومات قاعدة البيانات بنصوص من المدونة (Sinclair, 1996).

وتستعمل قواعد البيانات في توحيد المعلومات بإيكال توليد المعلومات المتشابهة إلى الحاسوب وهو ما قد التزمه مشروع معجم كوبيلد إذْ أوكِل إلى الحاسوب توليد «المطلع التعريفي» لكل معرف من المعرفات وهو ما أدى إلى إحكام لغة الشرح وضبط محتواه بناء على منهجية معتمدة في معالجة الشروح (Sinclair, 1996). كما يوكَل إلى الحاسوب أيضًا مهمة إجراء اختبارات لضبط الإحالات (Cross Reference) وضبط ما التزم به صانعو المعجم كما حدث في مشروع معجم لونجمان حيث أُنِيطَ بالحاسوب ضبط كلمات التعريف بحيث لا تخرج كلمات التعريف عن القائمة المعتمدة.

٤ ,١ - مراحل بناء المعجم الحاسوبي

- **التصميم**

تعد مرحلة تصميم المورِد المعجمي المرحلة الأولى في صناعة العمل المعجمي، فيها يحدد المعجميون طبيعة مورِدهم المعجمي وأهدافه؛ أهو معجم يستهدف المستعمل البشري أم معجم معد لأنظمة معالجة اللغات الطبيعية. ومنهج بنائه تحليلا وتمثيلا لمادته، إضافة إلى الشكل النهائي الحاسوبي الذين ينوون له أن يخرج فيه.

- **التحليل المعجمي الدلالي**
  - **التحليل الصرفي**

يمكن النظر إلى التحليل الصرفي في العمل المعجمي باعتباره مستويين:

الأول: هو مستوى التحليل الشكلي ويقصد به التحليل إلى جذر وساق، وإلى مجرد ومزيد. وهذا النوع من التحليل مهم في تحليل بنية المدونة النصية بغرض تخطيط الهيكلية العامة للعمل المعجمي المزمع إنتاجه.

الثاني: هو مستوى التحليل المعجمي الدلالي للكلمة؛ ويقصد به من وجهة نظر المعجمي تجلية علاقاتها الاشتقاقية الدلالية ببقية أفراد أسرتها الدلالية بما يحقق تمثيل البنية المعجمية الدلالية في أول مستوياتها؛ المستوى الصرفي. وتتجلى أهمية المستوى الصرفي في تحليل المحتوى المعجمي الدلالي للوحدة المعجمية فيما يلي:

- ضبط المحتوى الصرفي للوحدات المعجمية الاشتقاقية وتعميقه، وبالتالي ضبط طريقة تمثيلها في صورة تعريفات معجمية، عن طريق تنميط المعرّفات من الوجهة الصرفية، وهو أول خطوات ضبط لغة التعريف.
- تدقيق تذرية المحتوى الدلالي وتشريح طبقات المعنى، باعتبار الصرف هو الطبقة الأولى من طبقات الدلالة، وهو الأمر الذي سيكون له تأثيره في تعديد المعنى وتتبع تدرجه.
- ضبط العلاقات الاشتقاقية ببيان الأصل والفرع، وآليات الاشتقاق الدلالي، وبالتالي ضبط العلاقات الدلالية الموازية. وكل ذلك يؤدي إلى تعميق صياغة البنية الدلالية للمعجم وتجليتها من خلال التمثيل الدلالي؛ بضبط الاطرادات الصرفية الدلالية في المعجم لتحقيق الكفاية التفسيرية [٣٣].
- بناء قائمة المعجم البنية الكبرى (Macrostructure)، وبناء المدخل البنية الصغرى (Microstructure) وتنظيم معلوماته، وتوظيف العلاقات الاشتقاقية في تجلي البنية المعجمية، الدلالية.
- يعد الجانب الصرفي أحد معايير فصل المشتركات اللفظية (Homonyms) عن الوحدات المعجمية متعددة المعنى (polysemous). وبالتالي فإنه يوفر مادة تسهم في فك اللبس الدلالي آليًّا.

- **التحليل المعجمي**

ويقصد بالتحليل المعجمي تحليل الثروة اللفظية باعتبارها وحداتٍ معجميةً تبني قائمة مداخل المورِد المعجمي؛ ويكون ذلك بتحديد طبيعتها وكيفية إدراجها في المورِد المعجمي ومتطلبات معالجة محتواها المعجمي الدلالي. فالوحدات المعجمية متفاوتة من حيث طبيعتها وطبيعة محتواها المعجمي الدلالي. ومن مهام التحليل المعجمي تعيين الصيغة المعتمدة لكل مجموعة تنوعات صيغ (Paradigms) وهو ما يعرف بالتَّفريع أو تحديد رأس لمجموعة تنوعات شكلية لكلمة واحدة من أجل تحديد القائمة المعتمدة لكلمات أو مداخل المورِد المعجميّ.

وتتعاون مستويات التحليل (الصرفية والدلالية والتركيبية) في فحص جوانب الوحدات المعجمية، شكلا ومحتوى معجميا دلاليا. ويمكن توظيف نتائج التحليل في تنميط المعرفات كما يلي:

- أنماط المعرفات حسب المقولة المعجمية للوحدة المعجمية: وهو التصنيف المبني على التحليل المعجمي الذي يميز صنفين أساسين من الوحدات المعجمية، الوحدات المعجمية المفردة، والوحدات المعجمية متعددة الكلمات. وتُظهِر نتائج مرحلة التحليل المعجمي أنماط الوحدات المعجمية التالية:

- وحدة معجمية مفردة، وتشمل:

  - **الكلمة البسيطة**: وهي وحدة معجمية تامة لا يدخل في تكوينها وحدات أخرى.
  - **الكلمة المركبة أو المنحوتة**: وهي الوحدة المعجمية المصوغة – صرفيا – من أكثر من كلمة على سبيل النحت أو التركيب، ولكنها تعامل معاملة الكلمة على المستوى الشكلي (هجاء ونطقا)، وعلى المستوى التركيبي (فيكون لها قسم كلامي، وتأخذ مواقع تركيبية، وتكتسب حالات إعرابية)، والدلالي. ومن أمثلتها: بسمل: «قال بسم الله الرحمن الرحيم»، درعمي: «منسوب إلى دار العلوم».

- مختصر: وهو وحدة معجمية، يتم في الاستعمال اللغوي الاجتزاء بجزء منها عنها فتعاملها الموارِد المعجَمِيَّة معاملة المدخل المعجمي، وتدرجها في قائمة مداخلها. مثل: ا هـ بمعنى انتهى، ت بمعنى تليفون.
- **الاصطلاح (Idiom)**: وهو التعبير الذي يَظْهر في الاستعمال اللغوي مرتبط الأجزاء باعتباره من المسكوكات اللغوية، ولكن دلالاته لا يمكن توقعها من خلال معاني مفرداته؛ لكونه تركيبا سماعيا لا يمكن التعامل معه بتحليل مكوناته. ويمثل التعبير الاصطلاحي نمطا معجميًّا متميِّزًا بسبب هذه الخصائص المعجمية الدلالية.
- الاسم المركب: وهو عبارة عن تركيب من أكثر من كلمة على مستوى الشكل، يشير إلى مفهوم أو شيء مفرد من حيث المحتوى. وهو تركيب تتمتع مفرداته باستقلالية صرفية وتركيبية، ولكنه على المستوى الدلالي يُنْظر إليه باعتباره وحدة مستقلة. ويعد هذا الصنف نمطًا معجميًّا متميزا لطبيعة المحتوى التي تشبه طبيعة محتوى الوحدة المعجمية المفردة، وطبيعة الشكل الذي يشبه التركيب.

- وأما التصنيف بحسب المحتوى فباعتبار أن الوحدة المعجمية إما أن تكون وحدة معجمية لغوية، أو مصطلحية، أو موسوعية (Encyclopedic Unit).

وبالتمييز بين أنماط الوحدات المعجمية باعتبارها مداخل، يصبح المعجم مجموعة من القوائم المنمطة التي يمكن التعامل معها- حاسوبيا -باعتبارها ملفات، لكل منها متطلباتٌ لتمثيل محتواه المعجمي، وطريقة لتمثيله. وتظهر آثار التنميط في تمثيل المحتوى المعجمي الدلالي في الجوانب الآتية:

- تسهيل التعامل الحاسوبي مع المعارف اللغوية المتضمنة في المورِد المعجمي؛ لأن التعامل مع أنماط محددة، يسهل من تحليل النص المعجمي. كما يسهل عمليات الإحصاء المعجمي الآلي، وجعل نتائجها أدق وأصدق تعبيرًا عن الظواهر المعجمية.

- تنظيم العمل عند التحرير، وتحقيق الاقتصـاد أثناء عمليـــة التحرير، فالنمط الواحد من المعرفات يتم التعامل معه بطرق محددة يستعملها المحرر جاهزة.
- يؤدي إحكام النتائج في هذه المرحلة إلى ضبط تصميم قاعدة البيانات المعجمية؛ لكون التنميط أحد متطلبات إنشاء قاعدة البيانات المعجمية، التي يتم تصميم جداولها على أساس مخرجات هذه المرحلة من التحليل، وتحقيق تمثيل البنية ومادة المعجم.

- **التحليل التركيبي**

ويشمل التحليل التركيبي عدة أمور:

- تحديد أقسام الكلام: وهو تصنيف مبني على أسس تركيبية، إذ يعتمد مقولات أقسام الكلام معيارًا للتقسيم؛ نظرا للارتباط الوثيق بين المحتوى الدلالي والمحتوى التركيبي، مما يجعل من كل قسم من أقسام الكلام نمطا متميزا من الوحدات المعجمية لاختلاف المعالجة الدلالية ومتطلباتها بين أقسام الكلام المختلفة.

وتعد المقولة التركيبية أقدم المعايير التي تؤثر في تقنية التمثيل المختارة لمعالجة وحدات المعجم.

- تحليل البنية الحملية (Argument Structure) للوحدة المعجمية التي تظهر السلوك التركيبي المحتمل للوحدة المعجمية في الاستعمال اللغوي.
- تحليل بنية الحدث [1] الجِهِيَّة (Aspectual Event Structure) وهو جانب مهم في توصيف البنية التركيبية للفعل ومشتقاته. وتهتم التصورات النظرية المعجمية

١- ولعل أشهر تصنيفات بنية الحدث للفعل هي:
- الحالة وتكون في الفعل الذي يعبر عن صفة لازمة لصاحبها، مثل: جَبُنَ، جَحِدَ: قل خيره. و تعبر المعاجم العربية عن أفعال الحالة بعبارات منها: ما كان.
- نشاط مثل: جمع، جرَّب تجريبا. معالجة: مثل: جَذَّرَ العددَ: أخرج جذره، جبَر العظمَ: أصلحه.
- العمل مثل جلس، قام.
- تَحَوُّلٌ: ويكون في الفعل الذي يعبر عن انتقال الفاعل من حالة إلى حالة، مثل: أجدب المكان: صار مجدبا، تجَبَّنَ اللبن: صار جبنا.
- التحويل: مثل: جَعَلَ، أَجْلَسَ. و تعبر المعاجم العربية عن أفعال التحويل بعبارات منها: جعله، صيره.

الدلالية بتوضيح بنية الحدث الجهية باعتبار بنية الحدث بنية مركبة، تتكون من أحداث فرعية، وأن التوصل إلى هذه الأبنية الفرعية يساعد في توصيف الوحدة المعجمية دلاليا من جهة، كما يساعد في توصيفها تركيبيا بما يحدد سلوكها التركيبي من جهة أخرى.

وإذا كانت المعاجم التي تتوجه إلى المستعمل البشري لا تهتم بتوضيح نمط بنية الحدث فإن ذلك يعود إلى أن هذه المعاجم مقدمة للمستعمل البشري الذي يمكنه- بالسليقة- تركيب المفردات تركيبا متناغما دون تنافر دلالي تركيبي، للأنماط التي تأتلف والأنماط التي لا تأتلف. أما المعاجم الحاسوبية التي تجعل من بين أهدافها أن تكون موردًا للمعلومات المعجمية الدلالية للتطبيقات الحاسوبية فينبغي أن تُفَصَّل أنماط أبنية الحدث، على أنها لا تفصّل في ذلك التفصيل الموجود في التحليل الموجود في الدِّلالات المعجَمِيَّة ولكن فقط يشار إلى الحدث الأبرز في بنية الحدث. ويتوقف اعتماد المورِد المعجمي على أنماط دون غيرها على طبيعة مقاربته وأهداف مورِده المعجمي؛ وذلك كله في إطار التصور النظري الحاكم للعمل والموجه له.

- **التحليل الدلالي**

في مرحلة التحليل الدلالي يتم التعامل مع ظاهرة تعدد المعنى وما تستدعيه من قضايا أخرى مثل المجاز والاستعارة، وتمييز المشترك اللفظي عن متعدد المعنى. كما يشمل التحليل الدلالي الجوانب التالية: تصنيف الكلمة حسب حقلها الدلالي الذي تنتمي إليه. واكتشاف العلاقات الدلالية التي تقع الكلمة طرفا فيها.

إضافة إلى تعيين قيود الانتقاء (Selection Restrictions) للكلمات. إذ تمثل القيود الانتقائية أهمية كبيرة لأنظمة معالجة اللغات الطبيعية إذ يحاول المعجمي فيها محاكاة العقل البشري في قدرته التركيبية، التي تمكّنه من تعرُّف التراكيب مقبولة التأليف من تلك التي تعد غير مقبولة التأليف؛ لذا فإن الموارِد المعجَمِيَّة التي تستهدف أنظمة معالجة اللغات الطبيعية تعتني اعتناء كبيرا بتوضيح أنماط القيود الانتقائية للوحدات المعجمية.

▪ **التمثيل المعجمي**

يُعَدُّ التمثيل التحققَ الفعليَّ للمنهجية المعتمدة على المورِد المعجمي، كما أنه يعد الصياغة الرسمية لمخرجات مرحلة تحليل المدونة النصية. ويظل التعريف- حتى الآن- أهم أشكال تمثيل المحتوى المعجمي الدلالي، وأهم مصادر المعارف المعجمية على مستوى المستعمل البشري والتوظيف الحاسوبي.

ويختلف التمثيل للمستعمل البشري عن التمثيل للآلة: فالثاني أكثر عمقا وتفصيلا، وابتعادا عن الجوانب الضمنية التي تترك لسليقة المستعمل. ويهدف أيُّ تصورٍ يَرنُو إلى تقديم منهجية لتمثيل المحتوى المعجمي الدلالي لأنظمة معالجة اللغات الطبيعية إلى تحقيق مجموعة الأهداف التالية:

- التعميق المعجمي الدلالي للمحتوى، بحيث يقدم المعارف اللغوية اللازمة للتطبيقات الحاسوبية بصورة جلية.
- الصياغة المنضبطة للغة التمثيل المعجمي، وتمثيل البنية المعجمية الدلالية بالتوصل إلى البنية الذرية لجميع المستويات بما يتلاءم مع المقاربات الحاسوبية.
- التماسك والاتساق بين إجراءات التحرير فيما بينها من جهة، وبين المنطلقات النظرية للتصور النظري التي يتبناها المورِد المعجمي.
- الإسهام في فك اللبس، بتجميع الأشكال الممكنة للمفردة الواحدة، وتصريفاتها، وفصل المشتركات اللفظية، وفصل المعاني وتمييزها، وتوضيح القيود السياقية والقيود التركيبية، وتصنيف الوحدة المعجمية بحسب الحقل الدلالي، وتجلية العلاقات الدلالية، والسمات الدلالية.

ولعل أهـــم الإشكالات التي ينبغي أن يؤسس لها نظريا في مرحلة التمثيل، لدى أي محاولة لبناء مورِد معجمي هي منهجية تمثيل متعدد المعنى ومنهجية تمثيل البنية المعجمية.

- **تمثيل الوحدة المعجمية متعددة المعنى**

من القرارات الأولية ذات الأهمية اتخاذ موقف في طريقة التعامل مع متعدد المعنى، وتقنيات تمييز هذه المعاني المتعددة [١]. وطريقة تنظيم المعاني في متن المعجم أو قاعدة البيانات، فكما تتباين الموارِد المعجَمِيَّة في منهجيات تمثيل المحتوى المعجمي الدلالي، تتفاوت في وسائل تمييز معاني الوحدة المعجمية الواحدة متعددة المعنى. وتعد وسائل التمييز بين المعاني من أهم ملامح منهج التمثيل لأي مورِد معجمي؛ لما لها من تأثير في التطبيقات التي تقصد إلى فك الالتباس الدِّلاليّ.

وتتركز الإشكاليات التي يوليها المعجميون الاهتمام، لدى معالجة متعدد المعنى، في مستوى التمثيل المعجمي في جانبين:

- تقنيات التمييز بين المعاني: ويقصد بها تمييز المعجميين المعاني المتعددة للوحدة المعجمية. وتعود أهمية ذلك بالنسبة للعمل الحاسوبي إلى كونها من أكثر أسباب وقوع اللبس على المستوى الدلالي. حتى إنه لم يعد مقبولا من أي نظرية تتصدى للدلالة المعجمية عموما والدلالة المعجمية الحاسوبية على وجه الخصوص ألا تُقدِّمَ تصوراتها النظرية وإجراءاتها العملية لمجابهة تعدد المعنى تحليلا وتمييزا، بحيث يكون ذا خطوة واسعة في سبيل فك لبس الوحدة المعجمية في السياقات المختلفة التي من الممكن أن تقع فيها.
- تمثيل المعاني المتعددة وتحديد الروابط بينها: والمشكلة الثانية التي تفرض نفسها في مسألة تمثيل المعاني المتعددة هي مسألة تنظيم هذه المعاني وتحديد الروابط الدلالية فيما بينها. وهي قضية قديمة قدم الصناعة المعجمية، فهل تُسْرَد المعاني بلا أساس أم يعتمد أساس للترتيب وتوضيح العلائق البينية لهذه الدلالات. وفي ترتيب الدلالات تعددت الإستراتيجيات المقترحة والمنجزة في هذا المجال، يحكم اختيارها طبيعة المعجم وغايته، فتشمل هذه الإستراتيجيات التنظيم التاريخي، والمنطقي، والإحصائي الوصفي، والتفسيري. ومن الإشكالات

١- قدم بو سفينسين Bo Svensén مجموعة من محددات المعنى التي يتم تداولها في تحرير الموارِد المعجَمِيَّة، وهي: معايير صرفية، ومعايير سياقية Syntagmatic تركيبية، ومعايير استبدالية رأسية Paradigmatic، ومعايير برجماتِيَّة Pragmatic. (Spohr, 2012).

المتعلقة بمعالجة متعددة المعنى كيفية تمثيل المجاز والمجاز المرسل: أباعتباره دلالة للوحدة المعجمية، أم باعتباره دلالة لها خصوصية، ينبغي التعامل معها بطريقة مختلفة عن بقية الدلالات التي تعد المعاني الحقيقية.

- **تمثيل البنية المعجمية الدلالية**

البنية المعجمية الدلالية هي رؤية للمعجم ترى في جوانبه المعجمية الدلالية مقوماتٍ تُمَكِّنُ من صياغتها صياغةً تُجَلي أنماطَها والعلاقات التي تربط بينها، بحيث يتجلى المعجم في صورة منظومية. ولقد ازداد الاهتمام بالبحث في بنية المعجم بدخوله في السياق الحاسوبي إذ أصبح تحقيق البنية المعجمية مطلبا تقليديا أو أوليا من مطالب الحوسبة المعجمية. للدرجة التي دفعت بعض من يؤرخون لظهور مصطلح المعجمية الحاسوبية بظهور أطروحة آمسلر (Amsler) التي كان موضوعها فحص بنية تعريفات معجم ويبستر للجيب (Amsler, 1980).

أما البنية المعجمية في السياق الحاسوبي فقد فرضت على كل من تصدى لاقتراح منهجية شاملة لمعجم أن يقدم تصورا متكاملا لآليات هذه المنهجية وإجراءاتها لإظهار بنية المعجم محوسبةً. فقواعد البيانات المعجمية والأعمال الشبكية فرضت مفاهيمها البنائية النسقية على العمل المعجمي. وتختلف الموارِد المعجَمِيَّة فيما بينها في طريقة صياغة بنية معجمية دلالية للمعجم، فالمكانز المعجمية والأعمال الشبكية مثل: شبكة الكلمات وشبكة الأطر والشبكة الذهنية لهي أكثر إحكاما في صياغة البنية المعجمية الدلالية من المعاجم المقروءة آليا؛ وذلك نظرًا لطبيعة التمثيل المعجمي الدلالي الذي تقوم عليه، فتشييدها يقوم بالأساس على تمثيل العلاقات المعجمية الدلالية بين وحدات المعجم.

ويعد تمثيل النظريات المعجمية بنية المعجم الدلالية أحد المعايير التي يُعتمَد عليها في تقييم كفاية نظرية، وتفضيل تصور نظري على آخر منافس له. فالبنية المعجمية الدلالية ليست ترفا علميا بل مطلبا ضروريا لأية نظرية تتصدى لإنجاز تمثيل للمعجم.

٤ , ٢- التقييس المعجمي

يقصد بالتقييس المعجمي وضع مواصفات ينبغي تحقيقها في النص المعجمي على مستوى بنيته الكبرى وبنيته الصغرى، شكلا ومحتوى، وذلك لتنميط لغة التمثيل

المعجمي بحيث يسهل التعامل معها حاسوبيا عند إرادة معالجة اللغة آليًّا. ومن المشروعات التي تبنتها موارِد مُعجَمِيَّة هو مشروع تقييس المحتوى المعجمي وطريقة تقديمه في الموارِد الحاسوبية. ولعل مشروع (Lexical Markup Frame work - LMF) [1] هو النموذج الأشهر لتقييس الموارِد المعجميَّة. ويهدف المشروع إلى الاتفاق الموسع على طريقة في تمثيل المحتوى المعجمي الدلالي، حتى يتسنى الإفادة منها خصوصا في مجال معالجة اللغات الطبيعية.

٤, ٣- التقويم المعجمي

التقويم للموارِد المعجمية، وهو أمر له أهميته في تطوير العمل المعجمي عن طريق نقده وتقييمه باعتماد معايير للتقييم والمفاضلة بين المشروعات المختلفة. ومثل أي تصور ينبغي قياس درجة كفايته مقارنا بالتصورات المقترحة المنافسة (السابقة). والكفاية تعني مدى ما يحققه التصور النظري من دقة في تمثيل الظواهر اللغوية: ملاحظة ووصفا وتفسيرا. وقد تعددت معايير اختبار الفرضيات اللغوية لتغطي كل مجالات الظواهر المدروسة. فقد أعاد جاكندوف توظيف مقترح تشومسكي بخصوص مستويات الكفاية اللسانية للنظرية النحوية في مجال النظرية المعجمية، فاقترح جاكندوف ((Jackendoff, 1975) المستويات الثلاثة التالية: الكفاية الرَّصدِيَّة (Observational Adequacy) بأن يكون المورِد المعجمي ممثلا لبنية معجم اللغة التي يمثلها؛ وحداتٍ معجميةً واستعمالاتٍ. والكفاية الوصفية (Descriptive Adequacy) بأن تستطيع قائمة الوحدات المعجمية توصيف/ تمثيل البنية المعجمية الدلالية للمعجم: كيانات وعلاقات، وتمثيل مقولات التحليل المعجمي الدلالي، وأنماطه التي تم التوصل إليها. والكفاية التفسيرية (Explan-atory Adequacy) وتعني القدرة على تمثيل البنية المعجمية تمثيلا يوضح العلاقات، والاطرادات، والاختلافات، والفروق الدلالية. وأن تكون مصوغة صياغة بنيوية. وقد تعرضت هذه الأفكار للمراجعة والتطوير. ومما ينبغي ذكره في هذا السياق إضافة بوسطيوفسكي (Pustejovsky,1995) مستوى آخر، هو الكفاية التَّجريبيَّة (Empir-ical Adequacy) وتعني الصمود أمام تحقيق القدر الأكبر من النجاح في الاختبارات

١- رابط المشروع هو: http://www.lexicalmarkupframework.org.

الذي تتعرض له مادة المورِد المعجمي أثناء التعامل معها حاسوبيا عن طريق التطبيقات الحاسوبية (Ide & Romary, 2002).

وبصورة أكثر عملية يمكن تطوير مجموعتين من المعايير لقياس الكفاية اللسانية الحاسوبية للمورِد المعجمي:

- مجموعة المعايير التي تتوجه إلى طبيعة المحتوى المعجمي: وتشمل درجة استيعاب الوحدات المعجمية والمعاني/ الدلالات، ودرجة عمق التحليل اللساني للمعلومات المقدمة وجدواها في تطبيقات معالجة اللغات الطبيعية.
- مجموعة المعايير التــي تتوجــه إلى لغة تمثيل المحتـوى المعجمي: وتتمثـل في تحليلية التمثيل؛ أي قابلية المورِد المعجمي للتحليل النَّحـوِيّ لاستخـــلاص المعلومات والأنمـــاط اللغوية من المورِد المعجمي، وفاعليـــة تمثيـل البنية المعجمية.

٥- الموارِد المعجَمِيَّة العربية الحاسوبية

للغة العربية ثروة كبيرة من الموارِد المعجَمِيَّة التقليدية/ غير الحاسوبية، تتنوع مادةً ومنهجا وحجما، كما أن لها ثروة معجمية حاسوبية آخذة في النمو والتطور وفي الوقت ذاته استعان العمل المعجمي الحاسوبي بمادة الموارِد المعجَمِيَّة الورقية في بناء موارِدِه المعجمية.

٥, ١- أنماط الموارِد المعجَمِيَّة العربية في علاقتها بالعمل الحاسوبي

تم تطوير عدد من الموارِد المعجَمِيَّة العربية الحاسوبية على اختلاف في طبيعتها وغايتها ودرجة كفايتها. وفي ما يلي رصد لأنماط الموارِد المعجَمِيَّة العربية في علاقتها بالعمل اللساني الحاسوبي:

- ويكون بنقل مورِد معجمي تقليدي إلى الشكل الحاسوبي، باعتباره نصا. ويكون المورِد المعجمي في هذه الحالة مقروءا للمستخدم البشري باعتباره نسخة مرقمنة من معجم تقليدي، ويمكن للبرنامج الحاسوبي التعامل معها باعتباره نصا. وهذا المستوى هو أبسط أنماط التعامل الحاسوبي مع الموارِد المعجَمِيَّة.

- المعاجم التقليدية المحوسبة. ويكون بتخزين مادة الموارِد المعجَمِيَّة التقليدية في أشكال حاسوبية كقواعد البيانات؛ بما ييسر طريقة التعامل معها سواء للمستعمل البشري أو البرامج الحاسوبية. وهذه الموارِد المعجَمِيَّة هي، في الحقيقة، موارِد مُعجَمِيَّة تقليدية أُلبِست ثوبا حاسوبيا بإدخالها - عبر لوحة المفاتيح - كما هي ودون تغيير كبير في طبيعة المعالجة، باستثناء ما فرضته طبيعة الشكل الحاسوبي تنظيما للمادة وطريقة العرض والاستعلام، وتجعل هذه الأعمال علي أسطوانة أو أتيحت للبحث علي شبكة المعلومات الدولية (On-line Dic-tionary). وهذا النمط من الموارِد المعجَمِيَّة يمكن اعتباره معجما تقليديا رغم اعتماده على الحاسوب في أحد جوانبه، لما فيه من احتفاظ بكل خصائص المعجم التقليدي. ومن ذلك معجم الغني، والمعجم العربي الشامل، والمعجم الوسيط في نسخته المحوسبة وغيرها.
- موارِد مُعجَمِيَّة تقليدية استعانت في تنظيم مادتها وإحصائها وإخراجها بالحاسوب. فعلى مستوى الصناعة المعجمية بدأ توظيف الحاسوب في بناء عدد من الموارِد المعجَمِيَّة مثل معجم اللغة العربية المعاصرة.
- المورِد المعجمي الحاسوبي: ويقصد به بناء المورِد المعجمي على أسس من المفاهيم الحاسوبية خصوصا في مراحل تمثيل المعلومات المعجمية، بالاعتماد على تصور نظري؛ بما يحقق متطلبات المقاربة الحاسوبية للمادة المعجمية.

ومن أمثلة الموارِد المعجَمِيَّة العربية القائمة ما يلي:

- **شبكة الكلمات العربية**

شبكة الكلمات العربية هو مشروع منبثق عن المشروع المركزي (Global Word-Net)، ويقوم التصور الأساسي في هذا العمل على الاعتماد على المشروع الأساسي باللغة الإنجليزية وسحبه إلى العربية عبر معجم ثنائي اللغة: إنجليزي/ عربي. ويتبنى هذا العمل جل التصورات النظرية والأدوات التطبيقية لمشروع شبكة الكلمات الإنجليزية، بل إنه ينطلق من الإنجليزية متبنيا افتراضا يرى أن تلك المنهجية هو الطريقة المثلى لبناء مورِد معجمي عربي حاسوبي في أسرع وقت لتحقيق أكبر قدر ممكن من الإفادة في تطبيقات معالجة اللغة الطبيعية. ويمكن النظر إلى هذا المشروع باعتباره موردًا معجميا

ثنائي اللغة لا يغني بحال عن النهوض بمشروع لشبكة لفظية عربية تنطلق أساسا من المعجم العربي، ومن طبيعته الصرفية والتركيبية والدلالية.

- **مشروع المعجم العربي التفاعلي**

يهدف مشروع المعجم العربي التفاعلي إلى بناء معجم عربي حاسوبي (تفاعلي) ولعل هذا المشروع أكبر المشروعات المعجمية العربية لما توفر له من إمكانات فنية ومالية وسياسية وإعلامية، باعتباره مشروعا عربيا قوميا بات من ضرورات العصر المعلوماتي.

وقد دعا القائمون على المشروع عددا كبيرا من المختصين في هذا المجال لاستقصاء الجوانب الفنية والأفكار العلمية التي يمكن أن يفاد منها في بناء هذا المورِد المعجمي. وقدم هذا الحشد من (خبراء المعجم) عددا من الأوراق البحثية تكاد تغطي معظم جوانب بناء مورِد مُعجَمِيّ حاسوبي. ولكن يلاحظ في هذا السياق هو عدم تغطية الأوراق (ومن ثم التوصيات النهائية) أهم جوانب العمل المعجمي الحاسوبي وهو جانب تمثيل المحتوى المعجمي الدلالي. الذي لم تفرد له ورقة بحثية واحدة، ومن ثم اكتفي بالعمل بالاعتماد على المعاجم التقليدية في ذلك.

٥,٢- مراجعة نقدية للموارِد العربية المعجمية

حقق العمل المعجمي العربي الحاسوبي نتائج طيبة على المستويين؛ البحثي والتطبيقي، فعلى المستوى البحثي يوجد عدد من الدراسات التي تتركز على العمل المعجمي الحاسوبي من منطلقات لغوية أو من منطلقات حاسوبية. وعلى المستوى التطبيقي تم تطوير عدد من الموارِد المعجَمِيَّة العربية منها ما كان مبنيا على أسس معجمية دلالية مثل مشروع بروب-بنك العربية (Arabic PropBank) التي بنيت بالتوازي مع مجموعة من الموارِد المعجَمِيَّة للغات أخرى أهمها بروب-بنك الإنجليزية، واتخذت مدونتها من أعداد من جريدة النهار اللبنانية. وقد أخذت بعض المشروعات العربية تفيد من النظرية الدلالية، مثل بناء: قاعدة للدلالات المعجمية العربية مؤسسة على نظرية الحقول الدلالية، إذ تم فيها توظيف نظرية الحقول الدلالية في بناء مورِد مُعجَمِيّ دلالي تقوم بنيته على توظيف ٢٠ نمطا من أنماط العلاقات الدلالية، واعتمدت هذا المورِد المعجَميّ في مادته الأساسية على المكنز الكبير.

وبالرغم من النشاط الواضح في مجال العمل المعجمي العربي الحاسوبي فإن هناك ما يشبه الاتفاق بين العاملين في المجال المعجمي ومجال اللسانيات الحاسوبية على أن المواردِ المعجَمِيَّة العربية يعوزها الكثير من التطوير بحثيا وتطبيقيا لتواكب التطور الحادث لهذا المجال في السياق العالمي من جهة وليحقق متطلبات مجال اللسانيات الحاسوبية. ويمكن إيراد بعض الملاحظات على واقع العمل المعجمي العربي الحاسوبي كما يلي:

- مشكلات في التأسيس النظري: إذ تعتمد معظم قواعد البيانات العربية على مفهوم للحوسبة، يرى في نقل الموارِد المعجَمِيَّة التقليدية إلى جداول قاعدة البيانات، حوسبة تامة للمعجم، غير أن العمل الحاسوبي يقتضي القيام على أفكار حاسوبية، ابتداء من تصور الغايات التي من أجلها يُبنى المورِد المعجَميّ، ومنهجية التحليل والتمثيل، وتصور طبيعة المحتوى المعجمي الدلالي.

- فمن المفاهيم التي ينبغي أن يتخذ فيها موقف مبني علي درس نظري؛ الموقف من التجمعات اللفظية إذْ إنه غير واضح أو محدد وبه اختلاط، فالفروق ليست جليه بين أنواع هذه التجمعات. ويشتد الخلط عند الحديث عن التعبير الاصطلاحي Idiom. علي أن هذا الموقف ليس له أن ينضبط ما لم يعتمد علي تأسيسٍ نظريٍّ شامل، ومدونة محوسبة تحلل النتائج المستخرجة منها ثم يتم تمثيلها على أسس معجمية دلالية. إذ لا أمل في تعريف الحاسوب التَّعبيرات الاصطلاحية - مثلا - فيتعرف عليها آليًّا في النصوص التي تقدم إليه ويعرف معناها (مقابلها إذا كان الحديث عن تطبيقات الترجمة الآلية) ما لم تكن الأفكار النظرية واضحة في أذهان مطوريه ابتداءً.

- مركزية الصرف في المعالجات الحاسوبية للمعجم وطغيان فكرة (الجذر - الجذع - الوزن - المجرد المزيد ..) علي الفكر المعجمي الحاسوبي وذلك مردوده إلى أمرين: سهولة السيطرة علي مادة المعجم تنظيما واستعلاما من خلال الصرف (أو المحللات الصرفية) لصورية المقاربة الصرفية أو شكليتها وإمكان إخضاعها للحوسبة دون الدخول في غياهب الدلالة والتركيب والأمر الثاني هو بقايا تكبل بالفكر المعجمي التقليدي. علي أننا لا نعيب الاهتمام بالصرف إنما نعيب اعتداده المحور الوحيد الذي يدار عليه المعجم، وهذا موقف نظري أدى إلى إهمال البحث في جوانب بنية المعجم التي هي من أهم القضايا في مجال العمل

المعجمي الحاسوبي الغربي، على أن البنية المعجمية هي التي يتجلي فيها الزخم النظري الذي حكم المعجم تنظيرًا وتنظيمًا.

- عدم ظهور آثار الحاسوب في مجال العمل المعجمي الورقي أو الحاسوبي فلم نر مثلا معاجم للاستعمال أو معاجم التجمعات أو قوائم بأكثر الكلمات شُيوعًا ومعانيها. وهي الأمور التي ستكون معالجتها من خلال المدونة المحوسبة أمرًا سهلاً إن هي بنيت علي تنظير لساني. فإلي الآن لم يخرج لنا معجم بمواصفات معجم لونجمان للإنجليزية المعاصرة، أو كولينز كوبليد وكل ما رأيناه هي معاجم لا تقدم جديدا غير تسهيل عملية البحث، هذه المعاجم يمكن أن نطلق عليها المعاجم المحوسبة التقليدية.
- وأما ما يخص الجانب الحاسوبي فإن الحاسوبين قد أَوْلَوا الجانب الحاسوبي جل اهتمامهم على حساب الجانب اللساني واللغوي - وتلك مَهَمَّةُ اللسانيات الحاسوبية - فاكتفوا بمعالجة المعاجم الكائنة مع تطويعها للقالب الحاسوبي، دون محاولة اقتراح تصور نظري معجمي دلالي للمعجم الحاسوبي المنشود، فظلت التصورات الموجهة نحو المعجم الحاسوبي- رغم وجاهة كثير منها تطبيقيا- لصيقة بجدار الصرف (الاشتقاق والتصريف).
- التركيز على الجوانب السكونية للمعجم العربي بالتركيز على سرد الدلالات أو المعاني الخاصة بالكلمات، دون الجوانب الديناميكية التوليدية لرصد الآليات المتعلقة بآليات توليد الدلالات الجديدة في السياقات الجديدة.
- تقليدية المقاربة الحاسوبية للمعجم: فقد ورثت المعجمية العربية الحاسوبية جُل المشكلات النظرية عن المعجمية التقليدية لذا فإن معظم هنات المعجم التقليدي تظهر في المعجم الحاسوبي. إذ تتبنى المشروعات المعجمية الحاسوبية الأفكار المعجمية والتركيبية والدلالية التقليدية؛ فيظهر المعجم وكأنه نسخة من المعجم التقليدي اتخذت ثوبًا حاسوبيًّا، دون تبني تصور خاص في استكشاف البنية المعجمية الدلالية العربية. وتظهر التقليدية في مقولات التصنيف الصرفية والتركيبية والدلالية المعتمدة. واعتماد معظم الموارِد المعجَمِيَّة المحوسبة علي المعاجم التقليدية التي هي نفسها تفتقر إلى الأسس النظرية البحثية التي تجعلها مادة كافية بمطالب معالجة اللغات الطبيعية أو حتى أن تكون ذات كفاية وصفية

كموارِد مُعجَمِيَّة تتوجه للمستعمل البشري وهو ما لاحظه عبد القادر الفاسي الفهري على المعاجم العربية من قصور وافتقار للكفاية الوصفية ونقص في الاستيعاب وعدم النسقية أو الانتظام في جوانب النطق والصرف والتركيب والدلالة أو جانب التأصيل Etymology.

واعتماد مفاهيم لغوية تقليدية مما يؤثر على دقة النتائج التي يخرجها البحث الحاسوبي، وتحديث المفاهيم اللغوية والمعجمية ينبغي أن تسبق التحديث التقني إذ إن الأول هاد ومرشد للثاني لا العكس. ولا شك أن حل الإشكالات التنظيرية حلا نظريا في المعجم التقليدي ستظهر آثارها في المعجم الحاسوبي فالمعالجة الحاسوبية للمعجم مرآة تعكس المنجز التنظيري وتحصر نتائجها ثم تعيد تمثيلها، فلا يمكن أن نعتقد أن تقنيات التخزين ومعالجة المعلومات حاسوبيا ستكون معالجة منضبطة ما لم تكن قد توفرت لها ضوابط نظرية محددة قبلا، ثم يأتي العمل الحاسوبي تطبيقا لها.

- لا مجال لكل ما يقدمه الحاسوبيون من نقد للمعجم التقليدي؛ وإن كان جُلُّه صحيحا، فليس المطلوب أن يهجم الحاسوبي على مادة سائغة ثم يجعل الحاسوب يتعامل معها إنما ينبغي أن يوضع في الاعتبار أن المعجمية الحاسوبية هي الأخرى مطلوب منها صنع معاجم صناعة حقيقية ابتداء من التخطيط واتخاذ مدونة نصوصية ثم حوسبتها ووضع مخطط لمقولاتها الرئيسية والفرعية ثم التحليل المعجمي والتحرير فيكون أقرب إلى المعجم الذي نريده جميعا. على أن الحديث عن أي عمل معجمي حاسوبي دون أن تعد له العدة المعرفية اللسانية النظرية والتطبيقية الكافية، جنبا إلى جنب مع الأدوات الحاسوبية، لهو بمثابة قفز إلى النتائج دون معالجة المقدمات.

٦- الأفكار البحثية المقترحة في إطار العمل المعجمي الحاسوبي العربي

من واقع العمل المعجمي العربي الحاسوبي ومن خلال مقارنته بالواقع المعجمي الحاسوبي العربي يمكن استشراف آفاق العمل المعجمي الحاسوبي.

١- فهناك أعمال بحثية ودراسات ينبغي إنجازها مثل الأبحاث الدلالية المعجمية، والأبحاث في مجال الاستعمال، والبنية المعجمية، والبنية الاشتقاقية الدلالية

للمعجم العربي، ودراسات لسانية نفسية تقصد إلى توصيف المعجم الذهني لمتكلم العربية. مثل هذه الدراسات سوف تكون وسيلة لإنجاز مورِد مُعجَمِيّ عربي مؤسس على تأسيس نظري وتجريب عملي.

٢- ينبغي تطوير أدوات تحليل المدونات النصية العربية بتعميق جوانبها اللسانية حتى يمكن فرز التجمعات اللفظية والتَّعبيرات الاصطلاحية، بطريقة تتجاوز الجانب الإحصائي الغُفل الذي يمكّن من الحصول على التجمعات أيًّا كانت طبيعتها دون النظر إلى بنيتها الداخلية.

٣- ينبغي التأسيس (أو إنجاز البنية البحثية التحتية) لإنشاء موارِد مُعجَمِيَّة حاسوبية مختلفة لعل أهمها، كخطوة أولى، فيما يخص الواقع المعجمي الحالي؛ المعجم العربي المقروء آليًّا يضاهي معجم ويبستر السابع (Webster ٧) أو معجم لونجمان للإنجليزية المعاصرة، بحيث يكون مادةً للبحث والدرس المعجمي الحاسوبي، ومصدرًا لاستخلاص المعلومات التركيبية والدلالية، وتوليد موارِد مُعجَمِيَّة أخري منه أو بمساعدته مثل (Word net) أو (Arabic Frame Net).

٤- إنجاز عدد من الموارِد المعجَمِيَّة العربية الضرورية والتي لا يمكن إنجازها الآن إلا باعتماد آليات العمل المعجمي الحاسوبي نظرا لاحتياج المجتمع اللغوي والتطبيقات الحاسوبية إلى مادتها. ومن هذه الأنواع: المعجم التاريخي، والمعجم التأصيلي والمعجم الاستعمالي والمعاجم القطاعية: مثل معاجم التَّعبيرات الاصطلاحية والمتلازمات اللفظية، والأفعال العبارية .المعاجم ذات الأهداف التطبيقية الخاصة: مثل معاجم الترجمة الآلية.

٥- تطوير منصة لسانية حاسوبية لتقييس وتقييم الموارِد المعجَمِيَّة، بحيث يتم تصميمها لتحقق كلا من المتطلبات اللسانية والمتطلبات الحاسوبية، ولتكون مرجعية لتقييس وتقويم الموارِد المعجمية العربية.

ببليوجرافيا مرجعيَّة

١. ابن مراد (إبراهيم): مقدمة لنظرية المعجم، دار الغرب الإسلامي، بيروت ١٩٩٨.

٢. عمر (أحمد مختار) بمساعدة فريق عمل: معجم اللغة العربية المعاصرة، عالم الكتب، ٢٠٠٨.

٣. عمر (أحمد مختار) بمساعدة فريق عمل: المكنز الكبير: معجم شامل للمجالات والمترادفات والمتضادات، سطور، القاهرة، ط١، ٢٠٠٠.

٤. الفهري (عبد القادر الفاسي): المعجم العربي: نماذج تحليلية جديدة، توبقال للنشر، الدار البيضاء، ١٩٩٩.

٥. الفهري (عبد القادر الفاسي): المعجمة والتوسيط، المركز الثقافي العربي الدار البيضاء، ١٩٩٧.

٦. مدينة الملك عبد العزيز للعلوم والتقنية، والمنظمة العربية للتربية والثقافة والعلوم (ألكسو ALECSO): ورشة عمل معجم اللغة العربية التفاعلي:
http://www.almuajam.org/index.htm.

7. Ahlswede, T. & Evens, M. (1989). A lexicon for a medical expert system. In Relational models of the lexicon, Martha Evens (Ed.). Cambridge University Press, New York, NY, USA 97-111.

8. Amsler, R. A. (1980). The Structure of the Merriam-Webster Pocket Dic-tionary. Technical Report. University of Texas at Austin, Austin, TX, USA.

9. Atkins, B. S., & Rundell, M. (2008). The Oxford guide to practical lexicography. Oxford University Press.

10. Attia, M., Rashwan, M., Ragheb, A., Al-Badrashiny, M., Al-Basoumy, H., Abdou, S., A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields, Lecture Notes on Computer Science (LNCS): Advances in Natural Language Processing, Springer-Verlag Berlin Heidelberg, LNCS/

LNAI; Vol. No. 5221/2008; pp. 65-76 http://www.springerlink.com/content/100p13145723v162/ Aug. 2008.

11. Baker, C. F. & Fillmore, C. J. & Cronin, B. (2003). The Structure of the FrameNet Database. Int J Lexicography (2003) 16(3): 281-296 doi:10.1093/ijl/16.3.281.
12. Boas, H. C. (2009). Multilingual FrameNets in Computational Lexi-cography: Methods and Applications. Walter de Gruyter.
13. Boguraev, B. (Ed.). (1989). Computational Lexicography for Natural Language Processing. Longman Publishing Group, White Plains, NY, USA.
14. Byrd, R. J. (1986a). 'Dictionary Systems for Office Practice' in Pro-ceedings of the Grosseto Workshop 'On Automating the Lexicon', also available as IBM Research Report RC 11872.
15. Calzolari, N. (1989). The dictionary and the thesaurus can be combined. In Relational models of the lexicon, Martha Evens (Ed.). Cambridge University Press, New York, NY, USA 75-96.
16. Cheng-ming, G. & Huang, C. & Gong, J. & Li, J. (1994). The evolution of machine-tractable dictionaries. In Proceedings of the 15th conference on Computational linguistics - Volume 2 (COLING '94), Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, 1231-1234. http://dx.doi.org/10.3115/991250.991352.
17. Chodorow, M. S. & Byrd, R. J. & Heidorn, G. E. (1985). Extracting semantic hierarchies from a large on-line dictionary. In Proceedings of the 23rd annual meeting on Association for Computational Linguistics (ACL '85). Association for Computational Linguistics, Stroudsburg, PA, USA, 299-304. http://dx.doi.org/10.3115/981210.981247.
18. Clark, J. T. (2012). Lexicological Evolution and Conceptual Progress. HardPress.
19. Debenham, J. (2012). Knowledge Engineering: Unifying Knowledge Base and Database Design. Springer-Verlag New York Incorporated.

20. Dolk, D. R. (1988). Model management and structured modeling: the role of an information resource dictionary system. Commun. ACM 31, 6 (June 1988), 704-718.

21. Elkateb, S. & Black, W. & Vossen, P. & Rodríguez, H. & Pease, A. & Alkhalifa, M. & Fellbaum, C., Building a WordNet for Arabic. http://www.adampease.org/Articulate/publications/LREC.pdf.

22. Esuli, A. (2010). Automatic Generation of Lexical Resources for Opinion Mining. VDM Publishing.

23. Fellbaum, C. & Alkhalifa, M. & Black, W. & Elkateb, S. & Pease, A. & Rodríguez, H. & Vossen, P. (2006). Building a WordNet for Arabic. Proceedings of the the 5th Conference on Language Resources and Evaluation LREC2006, 2006. http://nlp.lsi.upc.edu/papers/fellbaum06.pdf.

24. Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cam-bridge, MA: MIT Press.

25. Fillmore, C. J. (2005). "Frame semantics". In: Brown, K. (ed.), En-cyclopedia of language and linguistics. Oxford: Elsevier.

26. Francopoulo, G. and Paroubek, P. (eds) (2013). Front Matter, in LMF Lexical Markup Framework, John Wiley & Sons, Inc., Hoboken, NJ USA. doi: 10.1002/9781118712696. fmatter.

27. Frawley, W. (1989). Relational models and metascience. In Relational models of the lexicon, Martha Evens (Ed.). Cambridge University Press, New York, NY, USA 335-372.

28. Granger, S. & Paquot, M. (2012). Electronic Lexicography. Oxford University Press.

29. Hartmann, R.R.K. (2001). Teaching and researching lexicography. (Ap-plied linguistics in action.) Edinburgh: Pearson Education.

30. Ide, N. & Véronis, J. (1994). Machine Readable Dictionaries: What have we learned, where do we go? Proceedings of the International Work-shop on the Future of Lexical Research, Beijing, China, 137-46.

31. Ide, N.& Romary, L. (2002). Standards for Language Re-sources Proceedings of the Third Language Resources and Evaluation Conference (LREC), Las Palmas, Canary Islands, Spain, 839-44.

32. Inkpen, D. (2009). Building a Lexical Knowledge-Base of Near-Synonym Differences. LAP Lambert Acad. Publ.

33. Jackendoff, R. (1975). Morphological and semantic regularities in the lexicon.Language 51. 639-671.

34. Jarmasz, M. & Szpakowicz, S. (2001a). Roget's Thesaurus as an Elec-tronic Lexical Knowledge Base. In NIE BEZ ZNACZENIA. Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji 40-lecia pracy naukowej. W. Gruszczynski, D. Kopcinska, eds., BialystokHalliday, M A K; et al 2004 Lexicology and corpus linguistics : an introduction. New York.

35. Karpova, O. & Kartashkova, F. (2009). Lexicology and terminology: a worldwide outlook. Cambridge Scholars.

36. Landau, S.I. (2001). Dictionaries: The art and craft of lexicography. (2nd ed.) Cambridge: Cambridge University Press.

37. Levin, B. (1993). English Verb Classes and Alternations. University of Chicago Press.

38. Litkowski, K. C. (2005). "Computational Lexicons and Dictionaries", Encyclopedia of Language and Linguistics (2nd ed.).Elsevier Publishers, Oxford.

39. Mel'ˇcuk, I. A. (1988). 'Semantic Description of Lexical Units in an Explanatory Combinatorial Dictionary: Basic Principles and Heuristic Criteria; in International Journal of Lexicography 1.3. 165–188.

40. Mel'ˇcuk, I. A. (1995). The Future of the Lexicon in Linguistic De-scription and the Explanatory Combinatorial Dictionary. In I.-H. Lee (ed.): Linguistics in the Morning Calm 3 (Selected Papers from SICOL-1992), Seoul, 181-270.

41. Mel'čuk, I.A. (1998). "Collocations and lexical functions". In: Cowie, A.P. (ed.), Phraseology: Theory, analysis and applications. Oxford: Clarendon Press. 23–54. Ogden, C.K. and I.A. Richards. 1923. The meaning of meaning. London: Routledge and Kegan Paul.

42. Diab, M. & Al-Badrashiny, M. & Aminian, M. & Attia, M. & Elfardy, H. & Habash, N. & Hawwari, A. (2014). Tharwa: A Large Scale Dialectal Arabic - Standard Arabic - English Lexicon. The 9th edition of the Language Resources and Evaluation (LREC) Conference, 26-31 May, Reykjavik, Iceland.

43. Oltramari, A. & Vossen, P. & Qin, L. & Hovy, E. (2013). New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems. Springer-Verlag GmbH.

44. Ovchinnikova, E. (2012). Integration of World Knowledge for Natural Language Understanding. Springer.

45. Palmer, M. & Gildea, D. & Kingsbury, P. "The Proposition Bank: An Annotated Corpus of Semantic Roles." Computational Linguistics, 31:1., pp. 71-105, March, 2005. http://verbs.colorado.edu/verb-index/.

46. Pustejovsky, J. & Boguraev, B. (1993). Lexical Knowledge Rep-resentation and Natural Language Processing, in Artificial Intelligence, http://dx.doi.org/10.1016/0004-3702(93)90017-6.

47. Pustejovsky, J. (1995). The Generative Lexicon, MIT Press.

48. Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard and Herbert Ernst Wiegand (Editors). Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Special Focus on Computational Lexicography. Berlin/New York: Mouton de Gruyter.

49. Russell, J. & Cohn, R. (2012). Framenet. Book on Demand.

50. Sinclair, J.M. (ed.) (1996). Looking Up: an Account of the COBUILD Project in Lexical Computing. London: Collins.

51. Spohr, D. (2012). Towards a Multifunctional Lexical Resource: Design and Implementation of a Graph-based Lexicon Model. Walter de Gruyter.

52. Svensén, B. (1993). Practical Lexicography: Principles and Methods of Dictionary-Making. Oxford University Press. Translated from the Swedish by J. Sykes and K. Schofield.

53. Vermon, L. (2012). Lexicology and Lexicography: Words and Ways. Webster's Digital Services.

54. Véronis, J. & Ide, N. (1991). An assessment of semantic information automatically extracted from machine readable dictionaries. In Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics (EACL '91). Association for Computational Linguistics, Stroudsburg, PA, USA.

55. Wilks, Y. & Fass, D. & Guo, C. & McDonald, J. & Plate, T. & Slator, B. (1988). "A Tractable Machine Dictionary as a Resource for Computational Semantics," in Bran Boguraev and Ted Briscoe (eds) Computational Lexicography for Natural Language Processing, Harlow, Essex, Longman.

56. Würzner, KK. (Hrsg.) & Pohl, E. (Hrsg.). (2012). Lexical resources in psycholinguistic research. Universitätsverlag Potsdam.

57. Zernik, U. (1991). Editor, Lexical acquisition: exploiting on-line re-sources to build a lexicon. Lawrence Erlbaum Associates, Hillsdale, NJ.

الفصل الثَّاني

المدَوَّنات اللُّغَوِيَّة

د. الـمُعتزّ بالله السَّعيد

١- في مفهوم المدَوَّنات اللُّغَوِيَّة.

٢- إرهاصات المنهج، وتطَوُّر دراسة المدَوَّنات اللُّغَوِيَّة.

٣- المدَوَّنات اللُّغَوِيَّة العربيَّة.

٤- أنواع المدَوَّنات اللُّغَوِيَّة.

٥- عَنونة المدَوَّنات اللُّغَوِيَّة.

٦- المدَوَّنات اللُّغَوِيَّة وآليَّة فهرسة النُّصُوص.

٧- مجالات الإفادة من المدَوَّنات اللُّغَوِيَّة.

٨- أفكارٌ بحثيَّة لأطرُوحاتٍ علميَّةٍ مُستَقبليَّة.

٩- مِن المواقع الإلكترونيَّة التَّعليميَّة والإرشاديَّة.

١- في مفهوم المُدَوَّنات اللُّغَوِيَّة

تُعنى لسانِيَّاتُ المدَوَّنة (Corpus Linguistics) بِالبحث في الظَّواهِر اللُّغوية وتفسيرِها من خلال مجموعةٍ مِن النُّصُوص التي تُمَثِّلُ الواقِعَ اللُّغَوِيّ؛ وهي ليست عِلمًا بالمفهوم الدَّقيق للعُلُوم؛ لكنَّها منهجٌ لُغَوِيٌّ حديثٌ نسبيًّا، يهدِف إلى وصف واقع اللُّغة اعتمادًا على مجموعة من النصوص الَّتي تُمَثِّل ذلك الواقعَ من خلال مناهج التَّحليل اللُّغَوِيّ (الوصفيّ والمِعياريّ والتَّاريخيّ والمقارن والتَّقابُلِيّ) ومُستوياته (الصَّوت، والبِنية، والتَّركيب، والدِّلالة، والمعجم)، كما يهدف إلى التَّحقُّق من فرضِيّاتٍ قائمةٍ حول لُغَةٍ مُعَيَّنة أو مجموعةٍ من اللُّغات المشتَركة في بعضِ خصائِصِها.

وأداةُ البحث في هذا المنهج هي «المدَوَّنة اللُّغَوِيَّة» (Linguistic Corpus) باعتبارِها مجموعةً مِن نُصُوص اللُّغة المكتوبة أو المنطوقة الَّتي يمكن التَّعاملُ معها آلِيًّا والتَّحَكُّمُ في بياناتِها ومُدخَلاتها بالإضافة أو الحذف أو التعديل من خلال قواعدِ بياناتٍ صُمِّمَت لتكونَ قادرةً على التَّعامُل مع هذه النُّصوص، حيثُ تُمَثِّلُ هذه القواعدُ مخزنًا كبيرًا للُّغة، يُرجَع إليه وقتَ الحاجة ويَتحمَّل أيَّ قدرٍ من النُّصوص الَّتي يُمكن أن تُضافَ إلى المادَّة الأساسيَّة للمدَوَّنة اللُّغَوِيَّة مُستَقبلاً.

ومادَّةُ المدَوَّنة اللُّغَوِيَّة ليست نُصوصًا تقييديَّةً أو عشوائية؛ لكنَّها كتلةٌ غيرُ منتظمةٍ من النُّصوص الَّتي تخضعُ لمجموعةٍ من الأسُس والمعايير، يُحَدِّدها الهدفُ المنشودُ من المدَوَّنة اللُّغَوِيَّة. فالمدَوَّنة الَّتي يُعتَمَد عليها في صناعة مُعجَم لُغَوِيٍّ تختلف مادَّتُها عن تلك المستَخدَمةِ في حَصر مجموعةٍ من الأنماط التركيبيَّةِ أو البِنَوِيَّةِ للُّغة؛ كما تختلِفُ مادَّةُ المدَوَّنةِ المستخدَمةِ في صناعة مُعجَمٍ تكراريٍّ عن تلك الَّتي يُعتَمَد عليها في صناعة المعجمات التَّاريخيَّة؛ وهكذا. كذلك فإنَّ المعالجَةَ الآليَّةَ للنُّصوص تتَّفِق وطبيعةَ المدَوَّنة؛ فالأدواتُ المستَخدمةُ وطريقةُ مُعالجَةِ النُّصوص وطرائقُ إدارةِ قواعدِ البيانات.. كلُّ هذا يخضَعُ لتلك الأسُس والمعاييرِ الَّتي تُحدِّدها طبيعةُ المدَوَّنة اللُّغَوِيَّة والهدفُ منها.

ومع وُضُوح الفكرة الرَّئيسةِ لاستخدام المدَوَّنات اللُّغَوِيَّة منذ ما يقرُب من أربعة قُرُون، إلاَّ أنَّ الطَّفرةَ المعلوماتيَّةَ الهائلةَ الَّتي شَهِدَها العصرُ الحديثُ في ميادين الحوسبة وتقنية المعلُومات قد غيَّرَت وُجْهَةَ الباحثين، ومَكَّنَتهم من التَّعامُلِ مع مجموعاتٍ ضخمةٍ من النُّصوص والحصولِ على نتائجَ أكثرَ دِقَّةً ووضوحًا. ولم يَعُد تفسيرُ الظَّواهِرِ

اللُّغَويَّة قاصِرًا على النَّظريَّات التقليديَّة؛ بل تخطَّى ذلك إلى التَّعامُل مع مناهجَ تجريبيَّةٍ أدَّت بدورِها إلى اختلافٍ في طرائق تحليلِ النُّصوص. كما لم يعُد بناءُ الأدوات والبرمجَيَّات المساعدة في مُعالَجَة اللُّغات الطَّبيعيَّة قاصِرًا على الخوارزمات، إذ أصبحَ لزامًا على صانِعيها أن يُفيدُوا من المدَوَّنات اللُّغَويَّة باعتبارها موردًا لُغَويًّا رئيسًا.

ويَخضَعُ اختِيارُ نُصُوص المدَوَّناتِ اللُّغَوِيَّةِ عندَ بنائِها لإحدى ثلاث طُرُق:

١ , ١ – **الطَّريقة الأولى**: تَقُومُ على **الاستبانة** (Questionnaire)، حيث يَطرَحُ صُنَّاعُ المدَوَّنة مجموعةً من الأسئِلة على أشخاصٍ يُمَثِّلُونَ المجتَمَعَ اللُّغَوِيّ الَّذي تنتمي إليه النُّصُوص. وتتعلَّقُ محاورُ الاستبانة بالحُقُولِ المعرِفِيَّة الَّتي تدعو الحاجةُ إلى طَرْقِها، وعناوين الكُتُب المختارَة، وأسماء الكُتَّاب والمصَنِّفين ذوي الاختصاص، وأوجُه المفاضَلَة بينَ هذه الأمور جميعا .وفي ضَوء نتائِج الاستبانة يُحَدَّدُ حجمُ المدَوَّنةِ اللُّغَوِيَّةِ المنشودة؛ وتُحَدَّدُ – كذلكَ – الحُقُولُ المعرِفِيَّةُ الَّتي تُصَنَّفُ إليها النُّصُوص، ومَصادِرُ المادَّة النَّصِّيَّة، والفَترَةُ الزَّمَنِيَّة الَّتي تنتمي إليها النُّصُوص، والمنطِقةُ الجُغرافِيَّةُ الَّتي ينتَمِي إليها الكُتَّابُ والمصَنِّفون .وتُستَخدَم هذه الطَّرِيقةُ – عادَةً – في بناءِ مُدَوَّنات الدِّراسات التَّجرِيبيَّة (Empirical Studies))، ومُدَوَّنات اللَّهجات [القديمة والمعاصِرة]؛ كما تُستَخدَمُ في بناءِ المدَوَّنات اللُّغَوِيَّةِ للمجتَمَعاتِ الإقليمِيَّة، بما يُساعدُ على صناعة الأطالس اللُّغَوِيَّة .وتجدُرُ الإشارةُ إلى شُيُوع هذه الطَّريقة في بناء المدَوَّنات المصنوعة لأغراضٍ تعليمِيَّة وتربويَّة، لاسيَّما عندَ بناء المعجَمات التَّعليميَّة والكُتُب الدِّراسيَّة الموَجِّهة إلى تعليم اللُّغات، سواءٌ أكانَت لأبناء اللُّغة أم للنَّاطِقِينَ بغيرها.

١ , ٢ – **الطَّريقة الثَّانِية**: تَقُومُ هذه الطَّرِيقة على **الحَصر الشَّامل** للنُّصُوص (Comprehensive Inventory)؛ وتُلزِمُ صُنَّاعَ المدَوَّنةِ بِحصرِ جميع نُصُوصِ المجتَمَع اللُّغَوِيّ الَّذي تُمَثِّلُه المدَوَّنة المنشودة، دونَ استثناء شيءٍ من مادَّة هذا المجتَمَع. وتُستَخدَمُ هذه الطَّرِيقةُ – عادَةً – عند بناءِ المدَوَّنات اللُّغَويَّة الَّتي تستهدفُ الدِّراسات المسحِيَّة (Survey Studies)؛ كما تُستَخدَمُ في بِناءِ المدَوَّنات اللُّغَوِيَّة ذاتِ المدى الزَّمَنِيّ أو الجُغرافِيّ المحدود، كالمدَوَّنات المستَخدَمة في صِناعَة مُعجَمات الأدباء، والمدَوَّنات المستَخدَمة في صِناعَة مُعجَمات الكتُب المقَدَّسَة، والمدَوَّنات المستَخدَمة في الدِّراسات الأدبيَّة – كما هُوَ الحالُ عندَ دراسة ظاهرةٍ أدبيَّةٍ مُعَيَّنَةٍ أو مجموعةٍ من الظَّواهر عندَ أديبٍ

مُعَيَّن، والمدَوَّنات المستخدَمة في الدِّراسات النَّفسيَّة والسُّلُوكيَّة - كما هُوَ الحالُ عندَ دراسةِ الظَّواهر النَّفسيَّة والسُّلُوكيَّة في مُجتمَعٍ ما.

١,٣- **الطَّريقة الثَّالِثة**: هي الأكثرُ شُيوعًا؛ وتَقــــُومُ على **نَظَرِيَّة العَيِّنات الإحصائِيَّة** (Statistical Sampling Theory)؛ ومِن خلالِها يقومُ صُنَّاعُ المدَوَّنةِ اللُّغَويَّة باختِيارِ عَيِّنةٍ مِن النُّصُوصِ الَّتي تَتَّفِقُ وأهدافَهم البَحثِيَّة، سواءٌ أكانَت عَيِّنَةً عشوائِيَّةً (Random Sam-ple)، أم غيرَ عشوائِيَّة (Non-Random Sample)، حيثُ يُنظَرُ إلى هذا النَّوع من المدَوَّنات باعتبارِهِ عَيِّنَةً مِن مُجتَمَعٍ غير محدُود، هو مُجتَمَع اللُّغة؛ كما يُلتَزَمُ عندَ بناء المدَوَّنات اللُّغَويَّة المصنوعة وفقًا لهذه الطَّريقة بأساليب التَّحليل الإحصائِيّ، بما يضمنُ أن تكونَ المدَوَّنةُ مُمَثِّلَةً لواقع اللُّغة ومُعَبِّرَةً عنه. ويَشيعُ استِخدامُ هذه الطَّريقةِ في صناعة المعجمات عُمومًا؛ لاسيَّما المعجَمات اللُّغَويَّة العامَّة، والمعجَمات اللُّغَويَّة التَّاريخيَّة، ومُعجَمات العُلُوم والفُنُون، والمعجَمات المصطَلَحِيَّة. وتُستَخدَمُ هذه الطَّريقةُ - كذلكَ - في الدِّراسات النَّحْوِيَّة، وصناعة أدوات المعالَجَة الآليَّة للُّغات الطَّبيعيَّة، وميادين حوسَبة اللُّغة الَّتي تُعنى باسترجاع المعلومات (Information Retrieval) والتَّرجمة الآليَّة (Machine Translation).

٢- إرهاصات المنهج، وتطَوُّر دراسة المدَوَّنات اللُّغَوِيَّة

أفادَ الهُنُودُ والصِّينِيُّونَ والعَرَبُ - قديمًا - من مجموعات النُّصُوص في بناء مُعجماتِهِم اللُّغَوِيَّة والتَّعَرُّف على دلالات الكلمات. وفي مرحلةٍ مُتأخِّرة من القُرُون الوُسطى كانت هناك بعضُ المحاولات الفردِيَّة - غير المنهجيَّة - لبناء المدَوَّنات اللُّغَوِيَّة والإفادة منها في فهرسة النُّصُوص وصناعة المعجمات والتَّقعيد النَّحويّ والدِّراسات التَّوراتِيَّة والأدَبِيَّة، بالإضافة إلى استخدامها في ميادين البحث اللُّغَوِيّ.

وكانت البدايةُ المؤرَّخ لها في مجال الدِّراسات التَّوراتيَّة، حيثُ قامَ النَّاشر الاسكتلنديّ «ألكسندر كرُودِن » (Alexander Cruden) (١٦٩٩-١٧٧٠) بجمع مادَّة الكتاب المقَدَّس باعتبارِها مُدَوَّنةً لُغَوِيَّة، واستخدَمَها في بناء فهارس ألفبائِيَّة لكلمات الكتاب المقَدَّس - بعهدَيه (القديم والجديد) - وما يتعلَّقُ به من موضوعات. وألحَقَ هذه الفهارسَ بالطَّبعة المفهرَسة الأولى من الكتاب المقَدَّس في عام ١٧٣٦م، ليصنعَ بذلكَ أوَّلَ فهرسٍ تفصيليٍّ يُعتَمَدُ في إنجازِهِ على مُدَوَّنةٍ لُغَوِيَّة.

وتلا ذلك استخدام المدَوَّنات اللُّغَوِيَّة في صناعة المعاجم؛ وكانت البداية من خلال مُدوَّنة الأديب الإنجليزِيّ «صموئِيل جونسون » (Samuel Johnson) (١٧٠٩-١٧٨٤) الَّتي أنجَزَها في عام ١٧٤٦م – بمُساعدة ستَّة من تلاميذِه؛ واستَمَدَّ مادَّةَ المدَوَّنة من الأعمالَ الأدبيَّة لويليام شكسبير (William Shakespeare) وجون ملتون (John Milton) وجون درايدِن (John Dryden) وغيرهم من أعلام الأدب الإنجليزيّ في ذلك الوقت.

صَنَع جونسون من مادَّة مُدَوَّنَتِه مُعجمًا كبيرًا لمفرَدات الإنجليزِيَّة، ووَسَمَه بـ «مُعجم اللُّغة الإنجليزِيَّة» (A Dictionary of the English Language). واشتَمَلَ المعجَمُ – الَّذي نُشِرَ كاملاً في عام ١٧٥٥م – على أكثرَ من أربعين ألف مدخل مُعجميّ، وما يزيدُ على مئةٍ وخمسين ألف تحليلٍ لُغَوِيٍّ لمفرَدات هذه الحُقُول، ليُصبحَ – بذلك – أضخمَ المعجمات اللُّغَوِيَّة للإنجليزيَّة وقتَ صُدُوره. ولا يزالُ هذا المعجمُ واحدًا من أهمِّ وأشمل مُعجَمات اللُّغة الإنجليزيَّة على الإطلاق.

واتَّسَعَ مجالُ استخدام المدَوَّنات اللُّغَوِيَّة في صناعة المعجم في القرن التَّاسع عشر الميلاديّ مع الحاجة إليها في بناء المعجمات اللُّغَوِيَّة التَّاريخيَّة؛ فاستُخدِمَت المدَوَّنةُ اللُّغَوِيَّةُ مادَّةً لمعجم اللُّغة الألمانيَّة (Deutsches WörterBuch) في عام ١٨٣٨م، ومُعجم اللُّغة الهُولنديَّة (Woordenboek der Nederlandsche Taal) في عام ١٨٤٩م، ومُعجَم الإنجلِيزِيَّة الحدِيثة (New English Dictionary) في عام ١٨٥٩م.

وفي مطلع القرن العشرين أمكنَ الإفادةُ من المدَوَّنات اللُّغَوِيَّة في تعليم اللُّغات، حيثُ قامَ عالِم النَّفس الأمريكيّ «إدوارد لي ثورنديك» (Edward Lee Thorndike) (١٨٧٤-١٩٤٩) ببناء مُدَوَّنةٍ لُغَوِيَّةٍ لاستخدامها في تعليم اللُّغة الإنجليزيَّة. وقامت فكرة ثورنديك على إعادة ترتيب مُفرَدات المدَوَّنة – الَّتي تربو على أربعة ملايين كلمة – بحسب أكثرها شُيُوعًا؛ وفي عام ١٩٢١م نشَرَ هذه المادَّة على هيئتها الجديدة في كتابه الَّذي وَجَّهَه للمعلِّمين بعُنوان (Teacher's Word Book, New York).

وأمكَنَ الإفادةُ من منهج ثورنديك في بناء مُدَوَّنَتِهِ–فيما بعد– في بناء ما يُعرَفُ بالمعجَمات التَّعليميَّة. ومن ناحيةٍ أخرى، فقد امتَدَّت هذه الفكرةُ من الاقتصار على

تعليم اللُّغة الأمِّ لأبنائِها إلى تعليم اللُّغات الأجنبيَّة للنَّاطِقِينَ بغيرها؛ كما ساهَمَت بصورةٍ مُباشرةٍ في تطوير مناهج علم اللُّغة التَّربويّ (Educational Linguistics).

واستُخدِمَت المَدَوَّنات اللُّغَويَّةُ -كذلك- في الدِّراسات النَّحوِيَّة قبلَ مُنتَصَف القرن العشرين، إذ فَطِنَ اللُّغَويُّونَ إلى أهمِّيَّتها في تمثيل واقع اللُّغة عندَ التَّقعيد لها. ففي عام ١٩٤٠م نشَرَ اللُّغوِيُّ الأمريكيّ «تشارلز فريز» (Charles Fries) (١٨٨٧-١٩٦٧) كتابَه «قواعد النَّحو الأنجلوأمريكيّ» (American English Grammar)، واعتمدَ فيه على مُدَوَّنةٍ لُغَوِيَّةٍ مجموعةٍ من الخطابات الرَّسميَّة لأعضاء الكونجرس الأمريكيّ. وبدا استخدامُ المَدَوَّنات اللُّغَوِيَّة في الدِّراسات النَّحوِيَّة أكثرَ منهجيَّةً ووُضُوحًا في «البحث المسحيّ لاستخدامات اللُّغة الإنجليزيَّة» (The Survey of English Usage) الَّذي أعدَّه الإنجليزيّ «راندولف كويرك» (Randolph Quirk) -بمُساعدة آخرين- بين عامَي ١٩٥٩م و١٩٦٨م، واعتمدَ فيه على مُدَوَّنةٍ لُغَوِيَّةٍ يصلُ عدد كلماتها إلى مليون كلمة.

ثُمَّ اتَّضَحَت ملامحُ المَدَوَّنات اللُّغَوِيَّة واكتَمَلَ منهجُ دراستها من خلال مُدَوَّنة جامعة براون القياسيَّة للأنجلوأمريكيَّة المعاصِرة (The Brown University Standar Corpus of Present-Day American English)، أو ما يُعرَف بـ «مُدَوَّنة براون» (Corpus Brown)، الَّتي أنجَزَها اللُّغَوِيَّان، التِّشيكيّ «هنري كوتشيرا» (Henry Kučera) (١٩٢٥-٢٠١٠) والأمريكيّ «نِلسون فرانسيس» (Nelson Francis) (١٩١٠-٢٠٠٢)، بتكليفٍ من جامعة براون في عام ١٩٦١م، لتكونَ أوَّلَ مُدَوَّنة لُغَوِيَّةٍ مُحوسَبة.

اشتَمَلَت مُدَوَّنة براون على أكثر من مليون كلمة، جُمِعَت من مصادر أمريكيَّة مُختلفة، وتنوَّعَت مادَّتُها بين الكُتُب والمقالات الصَّحفيَّة والوثائق الحُكُوميَّة والرِّوايات والقصص القصيرة والتَّقارير وغيرها. تحظى مُدَوَّنة براون بعناية اللُّغَوِيِّين والمعجَمِيِّين منذ ظُهورِها، إذ مهَّدَت الطَّريقَ لدراسة اللِّسانيَّات الحاسوبيَّة، كما مهَّدَت الطَّريقَ أمامَ العديد من المشروعات الكبرى في مجالات البحث اللُّغَوِيّ وصناعة المعجم.

Members/nns of/in the/at committee/nn include/vibe Mrs./np Milton/np Bernet/np ,/, Mrs./np J./np Clinton/np Bowman/np ,/, Mrs./np Rollie/np W./np Bradford/np ,/, Mrs./np Samuel/np Butler/np Jr./np ,/, Mrs./np Donald/np Carr/np Campbell/np ,/, Mrs./np Douglas/np Carruthers/np ,/, Mrs./np John/np C./np Davis/np 3/cd ,/, ,/, Mrs./np Cris/np Dobbins/np ,/, Mrs./np William/np E./np Glass/nn-tl ,/, Mrs./np Alfred/np Hicks/np 2/cd ,/, ,/, Mrs./np Donald/np Magarrell/np ,/, Mrs./np Willett/np Moore/np ,/, Mrs./np Myron/np Neusteter/np ,/, Mrs./np Richard/np Gibson/np Smith/np ,/, Mrs./np James/np S./np Sudier/np 2/cd ,/, and/cc Mrs./np Thomas/np Welborn/np ./.

الشَّكل٢ -١ : نموذج من مُدَوَّنة براون (Brown Corpus)[1].

وفي الفترة من ١٩٧٠ إلى ١٩٧٨م قامَ فريقٌ من الباحثين في جامعتَي لانكَستر وأوسلو - بالتَّعاوُن مع مركز الحوسبة النِّرويجيّ في مدينة بيرجن (Bergen) - ببناء «مُدَوَّنة لانكَستر-أوسلو-برجن» (Lancaster-Oslo-Bergen (LOB) Corpus) للُّغة الإنجليزيَّة، على غرار مُدَوَّنة براون من حيثُ منهج البناء وطريقة المعالَجة. واشتَمَلَت هذه المدَوَّنة على مليون كلمة إنجليزيَّة مكتوبة، مُوَزَّعةٍ على خَمسِمِئةِ مجموعةٍ بواقِع ألفيْ كلمةٍ لكل مجموعةٍ على حدة. وقد جُمِعَت مادَّةُ المدَوَّنة من الصُّحُف والمجلاَّت الإنجليزيَّة الَّتي نُشِرَت في المملكة المتَّحدة حتى عام ١٩٦١م.

٣- المُدَوَّنات اللُّغَوِيَّة العربيَّة

ظَهَرَ منهجُ دراسة المدَوَّنات اللُّغَوِيَّة المحَوسَبة في أمريكا وأوروبا في مطلع السِّتِّينِيَّات من القرن العِشرين. ومع هذا، فالمنهجُ لا يزالُ جديدًا على اللُّغة العربيَّة الَّتي لم تعرِف الطريقَ إليه إلاَّ قريبًا من القرن الحادي والعِشرين، مِن خلال مشروعاتٍ بحثِيَّة وأطروحاتٍ عِلمِيَّة معدودة، نعرِضُ لبعضِها فيما يلي:

٣,١- مُدَوَّنة «نايمَيخِن» (NIJMEGEN Corpus)

أنجَزَها فريقٌ بحثيٌّ في جامعة نايمَيخِن الهولندِيَّة في عامَيْ ١٩٩٥ و ١٩٩٦م، بإشراف المعجميّ الهولندي «يان هوخلاند» (Jan Hoogland)؛ وهي مُدَوَّنةٌ لُغَوِيَّةٌ مكتوبةٌ، جُمِعَت مادَّتُها من الصُّحُف والمجلاَّت والآداب العربيَّة، وتضُمُّ ما يزيدُ على مليُونَي كلمة. استُخدِمَت في صناعة مُعجمٍ لُغَوِيٍّ للعربيَّة والهولنديَّة.

1- https://github.com/lrscy/NLP-Assginments/blob/master/HW1/Problem4/brown/ca17.

٣,٢- المُدَوَّنة العربيَّة (Corpus Linguae Arabicae –CLARA)

أنجَزَها فريقٌ بحثيٌّ بمعهد دراسات الشرق الأدنى بجامعة تشارلز التِّشيكِيَّة في عام ١٩٩٧م؛ وهي مُدَوَّنةٌ لُغَوِيَّةٌ مكتوبةٌ، جُمِعَت مادَّتها من الدَّوريَّات العلمية والصُّحُف العربيَّة، وتضُمُّ خمسين مليون كلمة. تُستَخدَمُ هذه المادَّةُ لأغراض الصِّناعة المعجَمِيَّة [1].

٣,٣- المُدَوَّنة العربيَّة (LEUVEN Corpus)

أُنجِزَت بجامعة لوفان الكاثوليكِيَّة في بلجيكا بين عامَيْ ١٩٩٥ و ٢٠٠٤م؛ وتتنوَّع مادَّتها بين المكتوب والمنطوق؛ فالمادَّة المكتوبة مستقاةٌ من الصُّحُف والمجلات وكُتُب تعلُّم العربية وتضُمُّ ثلاثة ملايين كلمة، والمادَّة المنطوقة مُستَمَدَّةٌ من الإذاعات العربيَّة والمسرَحِيَّات وتشتمل على ٧٠٠ ألف كلمة. صُنِعَت هذه المدَوَّنةُ ليُستَفادَ منها في بناء مُعجمٍ عربيّ/ هولنديّ، يُلَبِّي حاجةَ مُتَعلِّمي العربيَّة من أبناء هولندا وبلجيكا.

وثمَّةَ بعضُ المدَوَّنات اللُّغَوِيَّة العربيَّة المتاحة لأغراض البحث العلميّ، نذكُر منها:

٣,٤- مُدَوَّنة Egypt

وضعَها مركزُ مُعالجَة اللُّغة والكلام في جامعة جون هوبكنز (John Hopkins) في عام ١٩٩٩. وهي مُدَوَّنةٌ لُغَوِيَّةٌ مكتوبةٌ ومُتوازية، مادَّتُها القُرآن الكريم وترجمة معانيه إلى الإنجليزِيَّة والفرنسِيَة، وتُصاحبُها بعضُ الإحصاءات الَّتي أُجرِيَت على نُصُوص القُرآن الكريم. تُستَخدَمُ هذه المدَوَّنة لأغراض التَّرجمة الآليَّة، وهي مُتاحةٌ بصورةٍ مجَّانِيَّة.

٣,٥- مُدَوَّنة العربيَّة المعاصِرة (Corpus of Contemporary Arabic)

وَضَعَتها الباحثة القَطَرِيَّة لطيفة السُّليطي ضِمنَ الأطروحة الَّتي تقدَّمت بها إلى جامعة ليدز للحُصُول على درجة الماجستير في عام ٢٠٠٤م، وعُنوانُها «تصميم وتطوير مُدَوَّنة لُغَوِيَّة للعربيَّة المعاصِرة» (Designing and Developing a Corpus of Contemporary Arabic). جُمِعَت نُصُوصُ المدَوَّنة من المجلاَّت وصفحات الويب، ويربو عدد كلماتها على ثمانمئة ألف كلمة؛ استَخدَمَتها الباحثةُ لأغراضٍ تعليمِيَّةٍ تتعلَّقُ بتعليم العربيَّة لغير النَّاطِقِين بها.

1- http://web.ff.cuni.cz/ustavy/usj/staré/veda/projekty/clara.htm.

٣,٦- مُدَوَّنة المعجم التَّاريخيّ للُّغة العربيَّة

وَضَعَها الباحثُ (المعتزّ بالله السَّعيد) ضِمنَ الأطروحة الَّتي تقدَّمَ بها إلى جامعة القاهرة للحُصُول على درجة الدُّكتوراه في عام ٢٠١١م، وعُنوانُها «مُدَوَّنة مُعجم تاريخيّ للُّغة العربيَّة: مُعالجَة لُغَوِيَّة حاسُوبِيَّة». جُمِعَت نُصُوصُ المدَوَّنة من التُّراث العربيّ المكتوب عبرَ العُصُور الأدبيَّة للعربيَّة بدءًا من عام ١٥٧م إلى وقت إنجازِها. ولهذه الـمُدوَّنة ثلاثة إصدارات، الأوَّلُ في عام ٢٠١١، ويربو عددُ الكلمات فيه على مئة مليون كلمة، والثَّاني في عام (٢٠١٤)، ويربو عددُ الكلمات فيه على مليار كلمة، والثَّالثُ في عام (٢٠١٨)، ويربو عددُ الكلمات فيه على مليار ونصف المليار كلمة. وجديرٌ بالذِّكر أنَّ هذه المُدوَّنة كانت الأساسَ الَّذي انطلَقَ منهُ مُعجمُ الدَّوحة التَّاريخيّ للُّغة العربيَّة؛ حيثُ قامَ الباحثُ ببناء مُدوَّنة الـمُعجم، ويُشرِفُ على تطويرِها.

٣,٧- المدَوَّنات اللُّغَوِيَّة لمؤَسَّسة «إل دي سي» (LDC Corpora)

أنجَزَت مُؤَسَّسة LDC-التَّابعة لجامعة بنسلفانيا الأمريكيَّة- العديدَ من الموارِد اللُّغَوِيَّة الَّتي تدعمُ العربيَّة المعاصِرة ولهجاتها الدَّارِجة (في مصر والشَّام والخليج العربيّ)؛ واعتمدَتَ أكثرُ هذه الموارِدِ على مُدَوَّناتٍ لُغَوِيَّةٍ، تُتيحُها المُؤسَّسة للباحِثين المعنِيِّين بحوسبة اللُّغة [بمُقابل]. ويعرِضُ الجدولُ التَّالي لبعض المدَوَّنات اللُّغَوِيَّة الَّتي تُتِيحُها مُؤَسَّسة LDC [1].

المدَوَّنة	مصدر البيانات	التَّطبيقات
جيجاوورد العربيَّة Arabic Gigaword	مكتوبة / وكالات الأنباء ٢٠٠٦، ٢٠٠٧، ٢٠٠٩، ٢٠١١	استرجاع المعلومات، ونمذجة اللُّغة، ومُعالجة اللُّغات الطَّبيعيَّة
البنك النَّحويّ العربيّ Arabic Treebank (أربعة أجزاء)	مكتوبة / وكالات الأنباء ٢٠٠٣، ٢٠٠٤، ٢٠٠٥، ٢٠٠٦، ٢٠١٠، ٢٠١١، ٢٠١٢، ٢٠١٤، ٢٠١٦	الاستخراج التِّلقائيّ للمحتوى، واسترجاع المعلومات لُغَوِيًّا، والكشف عن المعلومات

1- http://www.ldc.upenn.edu

المدَوَّنة الإخباريَّة المنطوقة للعربيَّة GALE Phase 4 Arabic Broadcast News Speech	منطوقة / البثّ الإخباريّ ٢٠١٨	التَّعرُّف على الكلام المنطوق
المدَوَّنة المُتوازية لمُنتدى العربيّ BOLT Arabic Discussion Forum Parallel Training Data	منطوقة / نقاشات علميَّة ٢٠١٩	التَّرجمة الآليَّة (الإنجليزيَّة والعربيَّة المعاصِرة)

الجدول ٢-١: من المُدوَّنات اللُّغويَّة لمؤسَّسة «LDC».

٤- أنواع المُدَوَّنات اللُّغَوِيَّة

يتحدَّدُ نوعُ المدَوَّنة اللُّغَوِيَّة وفقًا للهدف منها ومجالات الإفادة من نُصُوصِها. وثمَّةَ اعتباراتٌ لتصنيف المدَوَّنات، نُجملها على النَّحو التَّالي:

٤, ١- المُدَوَّنات اللُّغَوِيَّة باعتبار هيئة النُّصُوص. وتنقسم إلى:

- المدَوَّنات اللُّغَوِيَّة النَّصِّيَّة (Text Corpora)

وهي المدَوَّنات اللُّغَوِيَّة الَّتي تستمِدُّ مادَّتَها من مصادرَ مكتوبة، كالصُّحُف والمجلاَّت والكُتُب المطبوعة والوثائق المخطوطة والمنشورات والأطروحات العلميَّة. تعكس هذه المدَوَّنات واقعَ اللُّغة المكتوبة ويغلُبُ عليها أن تكونَ تمثيلاً لمستوى اللُّغة الفُصحى؛ تُستخدَم في الدِّراسات اللُّغَوِيَّة التَّاريخيَّة وبناء المعجمات والتَّنقيب في البيانات وبناء الأُنْطُولُوجيات وشبكات الكلمات، كما تُستخدَم في التَّعرُّف الآليّ على الكلام المكتوب (Optical Character Recognition -OCR). من أمثلتها: مُدَوَّنة أكسفُورد الإنجليزيَّة (Oxford English Corpus) الَّتي اكتمَلَ بناؤُها في عام ٢٠٠٦م؛ وتُستَخدَم مادَّتها - الَّتي تربو على مليارَيْ كلمة - في إنجاز الطَّبعة الثَّالثة من مُعجم أكسفورد [التَّارِيخِيّ] للُّغة الإنجليزِيَّة (Oxford English Dictionary).

- المدَوَّنات اللُّغَوِيَّة المنطوقة (Speech/Spoken Corpora)

هي المدَوَّنات اللُّغَوِيَّة الَّتي تستمِدُّ مادَّتَها من مصادرَ منطوقة، كالأفلام الوثائقيَّة والمسلسلات الإذاعيَّة ونشرات الأخبار والمحادثات الهاتفيَّة. تعكس هذه المدَوَّنات واقعَ اللُّغة المنطوقة ويغلُبُ عليها أن تكونَ تمثيلاً لمستوى اللُّغة الدَّارجة؛ تُستخدَم في

الدِّراسات اللُّغَوِيَّة الوصفِيَّة ودراسة اللَّهجات وبناء الأطالس اللُّغَوِيَّة والتَّعرُّف الآليّ على الكلام المنطوق (Automatic Speech Recognition –ASR). من أمثلتها: مُدَوَّنة «سانتا باربرا» للأنجلوأمريكيَّة (الإنجليزيَّة في الولايات المتَّحِدة) المنطوقة (Santa Barbara Corpus of Spoken American English) الَّتي أُنجِزَت في جامعة كاليفورنيا؛ وتُستَخدَم مادَّتها في دراسة اللَّهجات الأمريكيَّة.

٤, ٢- المُدَوَّنات اللُّغَوِيَّة باعتبار تعدُّد اللُّغة. وأنواعها:

- مُدَوَّنات أحاديَّة اللُّغة (Monolingual Corpora)

وهي المدَوَّنات اللُّغَوِيَّة الَّتي تستمِدُّ نُصُوصَها من لُغةٍ واحدة، ويغلُبُ عليها أن تُغَطِّيَ مُستوى لُغَوِيًّا مُعَيَّنًا (اللُّغة الفُصحى أو الدَّارجة)؛ تُستَخدَم في بناء المعجمات أحاديَّة اللُّغة، كما تُستَخدم في العديد من مجالات البحث في عُلُوم اللُّغة، مثل الإحصاء اللُّغَوِيّ والدِّراسات النَّحوِيَّة والدِّراسات اللُّغَوِيَّة الوصفِيَّة. من أمثلتها: «مُدَوَّنة كوبيلد» (CO-BUILD Corpus) الَّتي شارَكَ في تطويرِها فريقٌ بحثيٌّ مُشترك بين جامِعة برمنجهام ومؤسَّسة «كولينز» (Collins) للنَّشر؛ وتُستَخدَم مادَّتها – الَّتي تتجاوز ٤٥٠ مليون كلمة – في بناء وتطوير سلسلة المعاجم الإنجليزيَّة «كولينز – كوبيلد» (Collins Cobuild).

- مُدَوَّنات ثنائيَّة اللُّغة (Bilingual Corpora)

وهي المدَوَّنات الَّتي تستمِدُّ نُصُوصَها من لُغَتَين تنتميان إلى فصيلةٍ لُغَوِيَّةٍ واحدة أو فصيلَتَين؛ وتُستَخدَم في بناء المعجمات ثُنائِيَّة اللُّغة وتطبيقات التَّرجمة الآليَّة وتعليم اللُّغات. من أمثلتها: المدَوَّنة الثُّنائِيَّة للجُمَل بينَ العربيَّة والإنجليزيَّة (Sentence Aligned Bilingual Arabic English Corpus) الَّتي أنجزَتها شركة صَخر للإفادة منها في تقييم البرمجِيَّات وتطبيقات مُعالجَة اللُّغات الطَّبيعِيَّة الَّتي تقومُ بها الشَّركة، وتَضُمُّ ما يربو على مليون وثلاثمئة ألف جُملة باللُّغَتَين العَرَبيَّة والإنجليزيَّة.

- مُدَوَّنات مُتعدِّدة اللُّغات (Multilingual Corpora)

وهي المدَوَّنات اللُّغَوِيَّة الَّتي تستمِدُّ نُصُوصَها من عِدَّة لُغات؛ تُستَخدَم في أغراض المدَوَّنات ثنائيَّة اللُّغة على نطاقٍ واسع. من أمثلتها: المدَوَّنة مُتَعَدِّدة اللُّغات (Multilingual Corpus) الَّتي أعدَّها الباحث العراقيّ ستَّار الزّوينيّ ضمن أطروحَته للدُّكتوراه في جامعة

مانشستر، للإفادة منها في تطبيقات التَّرجمة الآليَّة؛ ونُصُوصُها مُستَمَدَّةٌ من ثلاث لُغاتٍ، هي الإنجليزيَّة (في سبعة ملايين كلمة) والسِّويديَّة (في مليونين وسَبعمئة ألف كلمة) والعربيَّة (في مليونين وخمسمئة ألف كلمة).

٤ ,٣- المُدَوَّنات اللُّغَوِيَّة باعتبار توافُق النُّصُوص. وتنقسم إلى:

- المَدَوَّنات اللُّغَوِيَّة المتوازية (Parallel Corpora)

وهي المَدَوَّنات اللُّغَوِيَّة الَّتي تستمِدُّ نُصُوصَها من لُغَتَين أو أكثر، وتكونُ النُّصُوص أصلاً في إحدى هذه اللُّغات - وتُسَمَّى (اللُّغة المصدر)، وترجمةً في اللُّغة [أو اللُّغات] الأخرى - وتُسَمَّى (اللُّغة [أو اللُّغات] الهدف)؛ من النَّماذج المَمَثِّلة للمدَوَّنات المتوازية: ترجمات العهد القديم (وتكون العبريَّة والآراميَّة فيها مصدرًا لغيرهما من اللُّغات)، وترجمات العهد الجديد (وتكون اليُونانيَّةُ فيها مصدرًا لغيرها من اللُّغات)، وترجمات القُرآن الكريم (وتكون العربيَّة فيها مصدرًا لغيرها من اللُّغات، كما يُعتَبَر (حجر رشيد) نموذجًا لهذا النَّوع من المَدَوَّنات حيثُ كُتِبَت نُصُوصُه بثلاث لُغاتٍ، هي المِصرِيَّة القديمة (لُغة الكَهَنة) والقِبطيَّة (لُغة الشَّعب) والإغريقيَّة (لُغة الحُكَّام).



وتتعدَّد وسائل الإفادة من المَدَوَّنات اللُّغَوِيَّة المتوازية حيث تُستَخدَمُ في بناء المعجمات ثُنائِيَّة اللُّغة وتطبيقات التَّعلُّم الآليّ والتَّرجمة الآليَّة، كما تُستَخدَم في تعليم اللُّغات والدِّراسات اللُّغَوِيَّة المقارنة (بين لُغات الفصيلة الواحدة، كالعربيَّة والعبريَّة) والتَّقابُلِيَّة (بين لُغات الفصائِل المتَعَدِّدة، كالعربيَّة والإنجليزيَّة)؛ وتحقيقًا للأهداف المنشودة من المَدَوَّنات المتوازية توضَعُ نُصُوصُها -جنبًا إلى جنبٍ- في قوالب متوازية، بحيث تظهرُ - في مصفوفاتٍ - كلمةً كلمة، أو جُملةً جملة، وهكذا. من أمثلتها: المَدَوَّنة المتوازية لوقائع البرلمان الأوربّي (European Parliament Proceedings Parallel Corpus) الَّتي أُنجِزَت خلال الفترة من ١٩٩٦م إلى ٢٠٠٩م، وتَضُمُّ نُصُوصًا مُتوازية بين الإنجليزيَّة وعشرين لُغة أخرى من لُغات الاتِّحاد الأورُوبيّ، هي (البلغاريَّة، والتِّشيكِيَّة، والدَّنمركيَّة، والألمانيَّة، واليونانيَّة، والإسبانيَّة، والإستونيَّة، والفنلنديَّة، والفرنسيَّة،

والمجَرِيَّة، والإيطاليَّة، واللِّيتوانيَّة، واللاَّتفيَّة، والهولنديَّة، والبولنديَّة، والبرتُغاليَّة، والرُّومانيَّة، والسِّلُوفاكيَّة، والسِّلُوفينيَّة، والسِّويديَّة).

ويقتَضِي منهجُ بناء المدَوَّنات اللُّغَوِيَّة المتَوازِية أن يلتزِمَ صُنَّاعُها ببعض الضَّوابط لتيسير مُعالجتها وفقًا للغرض الَّذي وُضِعَت المدَوَّنةُ لأجله. فبالإضافةِ إلى وُجوب الالتزام بمُحاذاةِ النُّصُوص، ينبغي ألاَّ يُتَصَرَّفَ في اللُّغة الهَدَف بما يُخالِفُ النَّصَّ الأصلِيَّ في اللُّغة المصدَر. وعلى سبيل المثال فإنَّنا نُتَرجم الجُملةَ الإنجليزيَّة (Obama Said that) إلى العربيَّة بالجُملة «قالَ أوباما ذلك»، ولا نقول «قالَ الرَّئِيس أوباما»؛ كما ينبغي ألاَّ يُغالَى في التَّرجمة الحرفِيَّة بما قد يُغَيِّرُ المعنى، فلا نقول – مثلاً – (Take the door) ترجمةً للجُملة «خُذ الباب»، وإنَّما نُترجمها بالجُملة (Close the door). ومن ناحيةٍ أخرى، ينبغي مُراعاةُ الجوانب البرجماتِيَّة/ التَّداوُلِيَّة Pragmatics الَّتي تتعلَّقُ باستعمال اللُّغة بين أهلِها – بما في ذلك ألوانُ الاستعارة والكناية والمجاز؛ ويُمكن التَّمثيلُ على هذه الجوانب بالجُملة (She is in the clouds) الَّتي تُتَرجَمُ حرفِيًّا إلى الجُملة العربيَّة «إنَّها في الغُيُوم»، بينما يُرادُ بها «إنَّها شاردة الذِّهن»؛ وتحقيقًا للهدف من المدَوَّنات المتَوازية، ينبغي تحديد هذه الجوانب بعنايةٍ وتمييزها في كلا اللُّغَتَين.

- <u>المدَوَّنات اللُّغَوِيَّة غير المتوازية (Non-parallel Corpora)</u>

وتُعرَف أيضًا بالمدَوَّنات المتقارِبة/ المتقابلة (Comparable Corpora)، وهي المدَوَّنات اللُّغَوِيَّة الَّتي تستمِدُّ نُصُوصَها من مجموعةٍ من اللَّهجات في لُغةٍ واحدة، أو لُغَتَين، أو مجموعة من اللُّغات، وتكونُ النُّصُوص أصلاً غيرَ مُتَرجمٍ في أيٍّ من اللُّغات الَّتي تَضُمُّها المدَوَّنة؛ ينْدُر استخدامُ هذا النَّوع من المدَوَّنات نظرًا لنُدرةِ النُّصُوص المتقاربة بين اللُّغات (كالعُقُود القانونيَّة بين الإنجليزيَّة والفرنسيَّة في مُقاطعات كندا والمناهج التَّعليميَّة بين العربيَّة والإنجليزيَّة في بعض دول المهجَر النَّاطقة بالإنجليزيَّة). تُستَخدَم المدَوَّنات المتَقارِبة في التَّعلُّم الآليّ وتطبيقات التَّرجمة الآليَّة، كما يُمكنُ الإفادةُ منها في تطبيقات فكّ الالتباس الدِّلاليّ للكلمات (Word Sense Disambiguation –WSD).

من أمثلتها: المدَوَّنة الدَّوليَّة للإنجليزيَّة (International Corpus of English)، الَّتي تهدف إلى وضع الفُرُوق الأساسيَّة بين لهجات اللُّغة الإنجليزيَّة في الدُّوَل النَّاطقة بها باستخدام مجموعةٍ من النُّصُوص المتقاربة بين لهجات هذه الدُّوَل.

٤ , ٤ – المُدَوَّنات اللُّغَوِيَّة باعتبار طبيعة النُّصُوص. وتنقسم إلى:

• المدَوَّنات اللُّغَوِيَّة المتخَصِّصة (Specialized Corpora)

هي المدَوَّنات الَّتي تَستمِدُّ نُصُوصَها من حقلٍ مُعَيَّنٍ أو مجموعةٍ مُعَيَّنةٍ من الحُقُول، سواءٌ أكانت حُقُولا معرِفيَّة – كالحُقُول العِلمِيَّة والقانُونيَّة والإخبارِيَّة، أم حُقُولاً تاريخيَّة – كحُقُول اللُّغة القديمة والوسيطة والمعاصِرة في الإنجليزيَّة وحُقُول العُصُور الأدبيَّة في العربيَّة، أم حُقُولا جُغرافِيَّةً كنُصُوص العربيَّة في وادي النِّيل والجزيرة العربيَّة وبلاد فارس؛ وقد يُقتَصَرُ فيها على النُّصُوص المكتوبة أو المنطوقة، أو تُجمَعُ من نُصُوص كاتبٍ أو أديبٍ مُعَيَّنٍ، كالنُّصُوص المسرحيَّة عند شكسبير، ونُصُوص الشِّعر عند طاغور الهنديّ؛ تُستَخدَم في الدِّراسات اللُّغَوِيَّة الوصفِيَّة ودراسة اللَّهجات والظَّواهر اللُّغَوِيَّة في لُغة الأدباء، وتُستَخدم – كذلك – في بناء الأطالس اللُّغَوِيَّة والمعجمات اللُّغَوِيَّة المتَخَصِّصة، مثل مُعجمات مُصطلحات العُلُوم ومُعجمات الأدباء. من أمثلتها: مُدَوَّنة لندن–لوند للإنجليزِيَّة المنطوقة (London–Lund Corpus of Spoken English –LLC) الَّتي أنجزَها اللُّغَوِيّ السِّويديّ «جان سفارتفيك» (Jan Svartvik) بتكليفٍ من جامعة لوند، وتستمدُّ مادَّتَها من اللُّغة الإنجليزِيَّة المحكِيَّة في لندن فيما يتجاوزُ نصفَ مليون كلمة.

• المدَوَّنات اللُّغَوِيَّة العامَّة (General Corpora)

وهي مُدَوَّنات لُغَوِيَّةٌ لا تَتَقَيَّدُ بنوعٍ مُعَيَّنٍ من النُّصُوص، بل تتنوَّعُ مادَّتُها بين مجموعاتٍ مُختلفةٍ من الحُقُول المعرِفيَّة والتَّاريخيَّة والجُغرافيَّة، وقد تجمعُ بين النُّصُوص المكتوبة والمنطوقة، أو مُستوياتٍ لُغَوِيَّةٍ مُتَعَدِّدة؛ تتعدَّدُ أغراضُ المدَوَّنات اللُّغَوِيَّة العامَّة، وتُستَخدَم في مُختلف ميادين البحث اللُّغَوِيّ وصناعة المعجم ومُعالَجة اللُّغات الطَّبيعيَّة، ويكثُر استخدامُها –تحديدًا– في تعليم اللُّغات والتَّرجمة الآليَّة. من أمثلتها: مُدَوَّنة مشروع «أُونْطُو–نُوتْس» (OntoNotes) الَّتي أنجَزَتها مُؤَسَّسة LDC في عدَّة إصداراتٍ بين عامَي ٢٠٠٧ و٢٠١١م باللُّغات العربيَّة والصِّينِيَّة والإنجليزِيَّة؛ وتتنوَّعُ مادَّتُها بين الحوارات المكتوبة ونشرات الأخبار.

٥- عنونة/ تذييل المدَوَّنات اللُّغَوِيَّة

عنونة/ تذييل المدَوَّنة اللُّغَوِيَّة (Corpus Annotation) عمليَّةٌ وسيطةٌ، تنتقلُ خلالَها النُّصُوصُ من صُورتها الأوَّليَّة (الخام) إلى صُورةٍ يسهُلُ التَّعامُلُ معها آليًّا (مُعَنونة)؛ ويُقصَدُ بعنونة نُصُوص المدَوَّنات اللُّغَوِيَّة إضافة معلوماتٍ توضِيحيَّة توصِيفِيَّة لكُلِّ وحدة لُغَوِيَّةٍ في النُّصُوص على حدة، بحيثُ تُصبِحُ المعلوماتُ مُلازِمةً للوحدات.

وتُمَهِّدُ هذه العمليَّةُ للمعالجَة الآليَّة لنُصُوص المدَوَّنات اللُّغَوِيَّة، كما تُساعد في إجراء الاختبارات اليدَويَّة للنُّصُوص، حيثُ تُقَلِّلُ من الجهد المبذول في مُتابعة وإحصاء توصِيفات المفرَدات؛ ومن ناحيةٍ أخرى يُمكِن الإفادة من العَنونات الملحَقة بمُدَوَّنةٍ لُغَوِيَّةٍ مُعَيَّنةٍ في عنونة مُدَوَّنة لُغَوِيَّةٍ أخرى في حال اتِّفاق المدَوَّنَتَين في أهدافِهما، واتِّفاقهما - كذلك - في أسلوب المعالجة الآليَّة لنُصُوصِهما.

وتختلفُ طريقة عنونة المدَوَّنات اللُّغَوِيَّة باختلاف أنواعها وأغراضِها البحثيَّة؛ بل قد تختلفُ طريقةُ العَنونة بين مُدَوَّنَتَين باختلاف المعلومات المستَخلَصَة منهما، وإن اتَّفَقَتا في النَّوع والهدف. فطريقة عَنونة المدَوَّنات اللُّغَوِيَّة المستخدمة في التَّحليل التَّركيبيّ - مثلاً - تختلفُ عن تلك المستخدمة في فكّ الالتباس الدِّلاليّ أو التَّحليل المعجميّ للنُّصُوص.

ويعمدُ بعضُ صُنَّاع المدَوَّنات اللُّغَوِيَّة المستخدمة في صناعة المعجم إلى الإبقاء على مُدَوَّناتهم في صُورتها الخام، حرصًا على تماسُك هيئة النَّصُوص عند الاستشهاد بها أو البحث فيها، بينما يعمدُ غيرُهم إلى وضع المدَوَّنات اللُّغَوِيَّة المعجميَّة في صُورَتَين، إحداهما خام (للاستشهاد والبحث) والأخرى مُعَنونة (للإحصاء والمعالجَة الآليَّة).

وتتَضَمَّنُ عَنونات المدَوَّنات اللُّغَوِيَّة معلوماتٍ توصِيفِيَّة لُغَوِيَّة (مثل: أقسام الكلام وأنماطها التَّركِيبيَّة والدِّلاليَّة)، وأخرى غير لُغَوِيَّة، تتمثَّلُ في المعلومات البيليوجرافيَّة، والخصائص الشَّكليَّة للنُّصُوص (مثل: نوع الخُطُوط وأحجامها - في المدَوَّنات اللُّغَوِيَّة النَّصِّيَّة، ومواضع النَّبر والتَّنغيم والمقاطع الصَّوتيَّة - في المدَوَّنات اللُّغَوِيَّة المنطوقة). وسنعرِضُ لبعض أنواع العَنونة في المدَوَّنات اللُّغَوِيَّة العربيَّة فيما يلي.

٥ , ١ – العَنونة التَّركيبيَّة (Syntactic Annotation)

وتُعنى بإضافة معلوماتٍ نحوِيَّةٍ – تركيبيَّة – إلى نُصُوص المدَوَّنة اللُّغَوِيَّة، نَتَبَيَّنُ من خلالها أقسامَ الكلام (Parts of Speech – "PoS") الَّذي تنتمي إليه المفرَدات، بحيثُ تُعنوَن كُلُّ مُفرَدةٍ برمزٍ لقسم الكلام الخاصّ بها؛ ويُمكن الإفادة من العَنونة التَّركيبيَّة للمدَوَّنات اللُّغَويَّة العربيَّة في تحليل المدَوَّنات المتَوازية وبناء المحلِّلات التَّركيبيَّة وحصر أنماط الجملة العربيَّة، كما يُستَفادُ منه في تطبيقات التَّرجمة الآليَّة والإحصاء اللُّغَوِيّ.

وتتمُّ العَنونة التَّركِيبيَّة للمدَوَّنات اللُّغَوِيَّة باستخدام إحدى وسيلَتَين:

- <u>العَنونة بتعيين أقسام الكلام (PoS Tagging)</u>

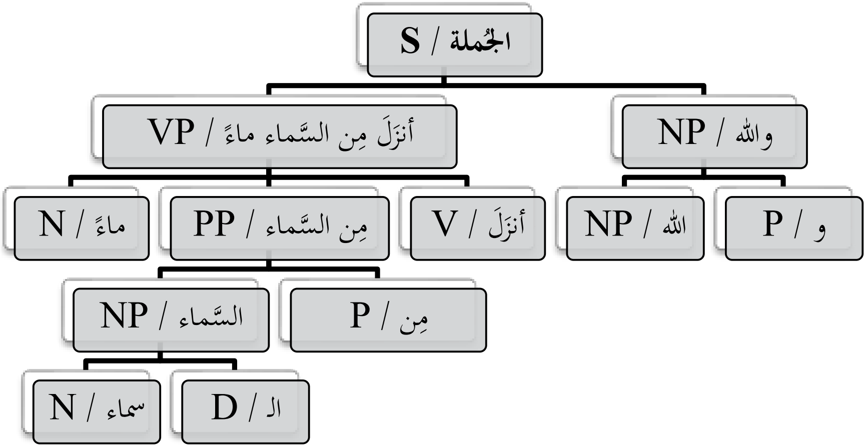
ويُقتَصَرُ فيها على إدراج قِسم الكلام لكلِّ مُفرَدةٍ دونَ النَّظَرِ إلى موقعها بين المفرَدات الأخرى في المدَوَّنة اللُّغَوِيَّة. ونُمَثِّلُ لها بالنَّموذَج التَّالي:

النَّصُّ قبلَ العَنوَنة
الآن.. ما الذي يشغل المثقفين المصريين أو العرب؟ ما هي القضية التي يختلفون حولها ويتفقون عليها؟
النَّصُّ بعدَ العَنوَنة
الآن RB / .. / ما RP / الذي WP / يشغل VBP / المثقفين NNS / المصريين NNS / أو CC / العرب NNS / ؟ / ما RP / هي PRP / القضية NN / التي WP / يختلفون VBP / حولها RB / ويتفقون VBP / عليها RP / ؟
رُمُوز عَنونة أقسام الكلام Tags PoS
الظَّرف RB، الأداة RP، الاسم الموصول WP، الفِعل المضارع VBP، الأسماء الشَّائعة (للجمع) NNS، حرف العطف CC، الضَّمير PRP، الأسماء الشَّائعة (للمفرَد) NN

الشَّكل ٢-٢: نموذج من مُدَوَّنة مُعَنونة بتعيين أقسام الكلام – من مقالات أنيس منصور.

- العَنونة بتحليل أقسام الكلام (Parsing)

وتُعنى بتحليل نُصوص المدَوَّنة اللُّغَوِيَّة إلى مجموعةٍ من الجُمَل، وتحليل الجُمَل إلى مُرَكَّبات/عبارات (Phrases)، وتحليل المَرَكَّبات إلى عناصِرِها الأوَّليَّة، ونستطيعُ التَّمثِيلَ لها بالنَّموذَج التَّالي:



الجُملة *Sentence*، المَرَكَّب الاسميّ *Noun Phrase*، المَرَكَّب الفِعليّ *Verb Phrase*، الاسم *Noun*، الفِعل *Verb*، شبه الجُملة من الجارّ والمجرور *Prepositional Phrase*، حرف الجرّ *Preposition*، أداة التَّعريف *Determiner*

الشَّكل ٢-٣: نموذج من مُدَوَّنة مُعَنوَنة بتحليل أقسام الكلام-من القُرآن الكريم (النَّحل: ٦٥).

وتجدُر الإشارةُ إلى تنوُّع الرُّموز المستَخدَمة في تعيين أقسام الكلام وتحليلِها بما يُحَقِّقُ الهدفَ المنشُودَ من المَدَوَّنة اللُّغَوِيَّة.

ونستَطِيعُ التَّمثِيلَ على ذلك بنَمَطَين من أنماط رُموز أقسام الكلام (PoS Tags)، الأوَّلُ هو ذلكَ النَّمَط الَّذي أقرَّته مُؤَسَّسة (Linguistic Data Consortium –LDC) بجامعة بنسلفانيا (Penn PoS Tags). ونُشِيرُ إلى بعض الرُّموز الَّتي يعتمِدُها هذا النَّمَطُ في الجدول التَّالي[1]:

1- LDC Website .(2011). List of Penn PoS tags used. From: https://catalog.ldc.upenn.edu/docs/LDC2003T06/arabic-POStags-collapse-to-PennPOStags.txt.

المصطلَح العربيّ	المصطلَح الإنجليزيّ	الرّمز
الصِّفَة	Adjective	JJ
الظَّرف	Adverb	RB
حَرف عَطف	Coordinating Conjunction	CC
اسم إشارة	Determiner / Demonstrative Pronoun	DT
كلمة أجنَبيَّة	Foreign Word	FW
نَكِرَة / شائِع (مُفرَد)	Common noun, Singular	NN
نَكِرَة / شائِع (جمع)	Common Noun, Plural	NNS
اسم عَلَم (مُفرَد)	Proper Noun, Singular	NNP
اسم عَلَم (جمع)	Proper Noun, Plural	NNPS
أداة	Particle	RP
فِعل مُضارِع / طَلَبِيّ	Imperfect Verb (***nb: imperfect rather than present tense)	VBP
فِعل مَبنِيّ للمجهول	Passive Verb (***nb: passive rather than past participle)	VBN
فِعل ماض	Perfect Verb (***nb: perfect rather than past tense)	VBD
أداة تَعَجُّب	Interjection	UH
ضَمِير شَخصِيّ	Personal Pronoun	PRP
ضَمِير مِلكِيَّة	Possessive Personal Pronoun	$PRP
عَدَد	Cardinal Number	CD
أداة رَبط / عطف	Subordinating Conjunction (FUNC_ WORD) or Preposition (PREP)	IN
اسم مَوصُول	Relative Pronoun	WP
ظَرف بِصِيغَة الاستِفهام	wh-Adverb	WRB

الجدول ٢-٢: من رُموز أقسام الكلام العربيّ **Penn PoS Tags** – عن «**LDC**».

والنَّمَطُ الآخَرُ أعدَّته الشَّركة الهندسِيَّة لتطوير النُّظُم الرَّقمِيَّة في مِصر RDI [عَطِيَّة & رشوان] - (RDI PoS Tags)، وهُو أكثرُ تفصِيلاً من سابِقِه، حيثُ يُعنى ببيان الحالة الصَّرفِيَّة للمفرَدات من حيثُ المصدرِيَّة أو الجُمود أو الاشتقاق، إلى جانب عنايته بأقسام الكلام. ونُشِيرُ إلى بعض الرُّموز الَّتي يعتمِدُها في الجدول التَّالي(1):

المصطلَح العربيّ	المصطلَح الإنجليزيّ	الرَّمز
اسم	Nominal	Noun
مَصدَر	Nouns made of infinitives	NounInfinit
اسم فاعِل	Subject noun	SubjNoun
صِيغة مُبالَغة	Exaggeration adjective	ExaggAdj
اسم مَفعُول	Object noun	ObjNoun
مُؤَنَّث	Feminine	Femin
مُذَكَّر	Masculine	Masc
مُفرَد	Singular	Single
مُثَنَّى	Binary	Binary
جَمع	Plural	Plural
حَرف جرّ	Preposition	Prepos
حرف نِداء	Interjection	Interj
اسم موصُول	Relative pronoun	RelPro
اسم إشارة	Demonstrative pronoun	DemoPro

الجدول ٢-٣: من رُموز أقسام الكلام العربيّ - "Attia & Rashwan" RDI PoS Tags.

1- Attia, M. & Rashwan, M., A Large-Scale Arabic POS Tagger Based on a Compact Arabic POS Tags Set, and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words, The Proceedings of the Arabic Language Technologies and Resources Int'l Conference; NEMLAR, Cairo-Egypt http://www.elda.org/nemlar-conf, Sept. 2004.

٥,٢- العَنونة الدِّلاليَّة (Semantic Annotation)

تختصُّ العنونة الدِّلاليَّة بالمفردات ذات الدِّلالات المتعَدِّدة في المدَوَّنة اللُّغَوِيَّة، ويُستخدَم في العديد من تطبيقات الذَّكاء الاصطناعيّ (Artificial Intelligence -AI) ومُعالجَة اللُّغات الطَّبيعيَّة (Natural Language Processing -NLP)، لاسيَّما تلك الَّتي ترتبطُ بالتَّحليل الدِّلاليّ، مثل شبكات الكلمات (WordNets) والشَّبكات الدِّلاليَّة (Semantic Nets) والأُنْطُولُوجْيات (Ontologies) وغيرِها؛ أضِف إلى ذلك أهمِّيَّته في إعداد المدَوَّنات اللُّغَوِيَّة المعجميَّة، وتصنيف النُّصُوص (Text Classification).

وتقومُ فكرة العَنونة الدِّلاليَّة على التَّمييز بين دلالات كُلِّ مُفرَدةٍ على حدة باستخدام خوارزمات فكّ الالتِباس الدِّلاليّ (WSD Algorithms)؛ وتجدُر الإشارةُ إلى أنَّ المدَوَّنات اللُّغَوِيَّة العربيَّة تتطلَّبُ عنونةً دلاليَّةً واسعَة النِّطاق إذا خلَت نُصُوصُها من علامات ضبط الحُرُوف (التَّشكيل). فالكلمة (بل) - على سبيل المثال - تحمل دلالات أقسام الكلام الثَّلاثة «الاسم (بَلٌّ) والفِعل (بَلَّ) والحرف (بَلْ)» عندَ خُلُوِّها من علامات الضَّبط، بينما لا تحملُ إلاَّ دلالات أحد أقسام الكلام عندَ ضبطِها بالشَّكل. وثمَّةَ طريقتان للعنونة الدِّلاليَّة الَّتي تُستَخدَم في النُّصُوص العربيَّة، تعتمد الأولى منهما على تمييز دلالة الكلمة في سياقِها، على نحو ما يُبيِّنُ الجدولُ التَّالي:

السِّياق			الدِّلالة
فرفعَ البلقيطيّ	**حاجبيه**	الخفيفين .. وقال: .. أنعِم وأكرِم	١
وتساءَلَ وهُو يَنتِفُ	**حاجبه**	الأيسر متى يمتلئ جيبي بنقود الحُكومة	١
فدخَلَ عليه	**حاجبه**	فأخبرَه أنَّ رَجُلا من الخوارج جيءَ به	٢
فلما طالَ انتظارُه، قالَ	**للحاجب**	إن أذِنتَ لي عليه، وإلاّ هجَوتُ اليَمَن .	٢
وقالَ جمال الدِّين بن	**الحاجب**	ابن عبد السَّلام أفقهُ من الغزالي	٣
ومِن الخُطَباء عُطارد بن	**حاجب**	بن زرارة، وقد خطَبَ أمام الرَّسُول	٣
(١) الشَّعْر النَّابِت فَوقَ العَيْن، (٢) خازن الباب وحارِسه، (٣) عَلَم / من أسماء الذُّكور			

الجدول ٢-٤: نموذج من مُدَوَّنة مُعَنونة دلاليًّا - طريقة١.

أمَّا الطَّريقةُ الأخرى، فتبدو أكثرَ موافقةً لطبيعة اللُّغة العربيَّة الاشتِقاقِيَّة، مع ما تتطلَّبُه من وقتٍ وجهدٍ لإنجازها على الوجه المنشود؛ ومفادُ هذه الطَّريقة أن يُرمَزَ إلى ثلاثة جوانب رئيسةٍ، هي: قسم الكلام الَّذي تنتمي إليه المفرَدة، ودلالة المفرَدة، وموضِع المفرَدة في النَّصّ الَّذي وَرَدَت فيه؛ ونستَطِيعُ - من خلال هذه الجوانب الثَّلاثة - أن نُمَيِّزَ بينَ دلالات المفرَدات في النَّصّ باعتبارِ أقسام الكلام PoS، وهو ما يعني ضرورةَ إخضاع النُّصُوص للعنونة التَّركيبيَّة بتعيين أقسام الكلام في مرحلةٍ أولى، ثُمَّ عنونتها بدلالة المفردة في مرحلةٍ ثانية، والنَّصّ على موضِع المفرَدة في سياقِها في مرحلةٍ ثالثة.

ولتوضِيح هذه الطَّريقة، نعرِضُ في الجدول التَّالي نموذَجًا للعنونة الدِّلاليَّة للمجموع الكتابيّ (من) في بعض سياقاته الَّتي وَرَدَت في القُرآن الكريم:

(م / د / PoS)	السِّياق (من: القُرآن الكريم)	PoS
(.. / ١ / ١)	﴿وَظَلَّلْنَا عَلَيْكُمُ الْغَمَامَ وَأَنزَلْنَا عَلَيْكُمُ الْمَنَّ وَالسَّلْوَى﴾	N
(.. / ٢ / ١)	﴿يَا أَيُّهَا الَّذِينَ آمَنُواْ لاَ تُبْطِلُواْ صَدَقَاتِكُم بِالْمَنِّ﴾	N
(.. / ٣ / ١)	﴿قَالُوا يَا وَيْلَنَا مَن بَعَثَنَا مِن مَّرْقَدِنَا﴾	N
(.. / ٤ / ٢)	﴿وَنُرِيدُ أَن نَّمُنَّ عَلَى الَّذِينَ اسْتُضْعِفُوا فِي الْأَرْضِ﴾	V
(.. / ٢ / ٢)	﴿يَمُنُّونَ عَلَيْكَ أَنْ أَسْلموا قُل لَّا تَمُنُّوا عَلَيَّ إِسْلَامَكُم﴾	V
(.. / ٥ / ٣)	﴿فَتَلَقَّى آدَمُ مِن رَّبِّهِ كَلِمَاتٍ فَتَابَ عَلَيْهِ﴾	P
المعاني		
(١) طَلٌّ ينزل مِن السَّماء	(٢) مَنَّ: تَفاخَرَ بالأنعام	(٣) اسم استفهام
(٤) أَنْعَمَ	(٥) حَرف جَرّ	
(م) = موضِع المفرَدة في النَّصّ/ السَّطر، (د) = دِلالة المفرَدة، (PoS) = أقسام الكلام N = 1 = الاسم، V = 2 = الفِعل، P = 3 = الأداة		

الجدول ٢-٥: نموذج من مُدَوَّنة مُعَنونة دلاليًّا - طريقة ٢.

٥,٣- التَّرميز (Encoding)

يُعنى التَّرمِيزُ بإضافة معلوماتٍ توصِيفِيَّةٍ لهيكل المدَوَّنة اللُّغَوِيَّة في صُورةٍ تُمَكِّنُ من التَّعامُل معها برمجِيًّا أو تحليلِ نُصُوصِها باستِخدام أدوات المعالجَة الآلِيَّة؛ ويُستَفادُ من هذه المعلوماتِ في بناء قواعد بيانات المدَوَّنات اللُّغَوِيَّة بصورةٍ مُنتَظِمة تُساعدُ في استرجاع النُّصُوص وقتَ الحاجة إليها؛ ومِن ناحِيةٍ أخرى يُساعِدُ التَّرميزُ فِي التَّعامُل مع المدَوَّنة اللُّغَوِيَّة وأدوات المعالجَة الآلِيَّة لها عبرَ الشَّبَكة العنكبوتِيَّة من خلال استِدعاء بيانات الويب (Web Mining)، كما يُمَكِّن من التَّعامُل المباشِر مع تِقنِيات التَّنقيب في البيانات (Data Mining)، وما يَتَفرَّعُ عنها، كالتَّنقيب في النُّصُوص (Text Mining) والبحث في مُستودعات البيانات (Data-Warehouse).

وتُستَخدَمُ لضبط هيئة النُّصُوص - تمهيدًا لترميزِها- إحدى صِيَغ النِّظام الموَحَّد لشفرات الحُرُوف (Unicode Transformation Format -UTF)، حيثُ تتوافَقُ هذه الصِّيَغ مع المعايير القِياسِيَّة لإظهار المحارِف أو الجرافيمات (Graphemes)، كما تدعَمُ العدِيدَ من ألفبائِيَّات اللُّغات الطَّبيعِيَّة، وإن كانَ يَعيبُها كِبَرُ المِساحة التَّخزِينِيَّة الَّتي تشغلُها المحارِف. أمَّا ترمِيزُ المدَوَّنات اللُّغَوِيَّة فيتمُّ باستخدام لُغة التَّوصِيف القابلة للامتِداد (Extensible Markup Language -XML)؛ وهي لُغَةٌ مُفَصَّلةٌ لبيانات المدَوَّنة اللُّغَوِيَّة، تدعَم نِظامَ الحُرُوف الدَّوْلِيّ الموَحَّد (Unicode)، وتَعمَلُ كقاعِدة بياناتٍ يَسهُل تناقلُها عبرَ صَفحات الوِيب.

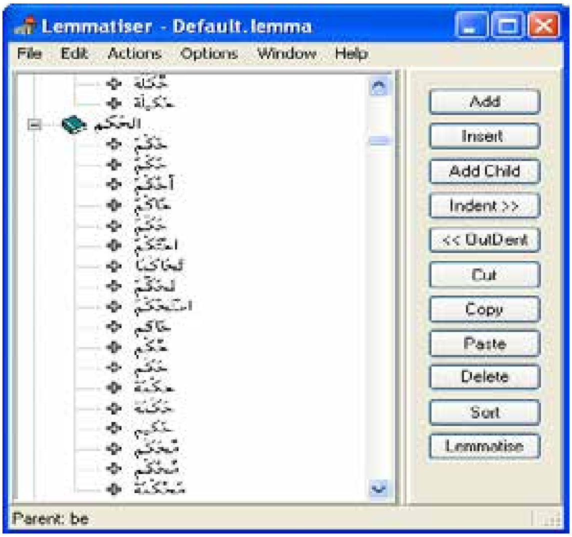
```
<?xml version='1.0' encoding='UTF-8' standalone='yes'?>
    <o:Title>اللائحة الأساسية لمجلس النواب المصري</o:Title>
    <o:Author>المُشَرِّع</o:Author>
    <o:Created>2009-08-14T12:18:00Z</o:Created>
    <o:LastSaved>2009-08-14T12:18:00Z</o:LastSaved>
    <resp>compiled by</resp>    <o:Words>1565</o:Words>
    <o:Characters>8921</o:Characters>
    <o:Lines>74</o:Lines>
    <name>Almo3taz Bellah</name>
    <w:t>اللائحة الأساسية لمجلس النواب المصري</w:t>
    <w:t>اللائحة الأساسية التي وافق عليها مجلس النواب المصري وصدر بها الأمر العالي</w:t>
    <w:t>(في 18 ربيع الأول سنة 1299 هـ)  7 فبراير سنة 1882</w:t>
    <w:t>نحن خديو مصر</w:t>
    <w:t>بعد الاطلاع على أمرنا الصادر بتاريخ 11 ذي القعدة سنة 1298 الموافق 4 أكتوبر سنة 1881.</w:t>
    <w:t>وبناء على ما قرره مجلس النواب، وموافقة رأي مجلس نظارنا.</w:t>
```

الشَّكل ٢-٤: نموذج من مُدَوَّنة مُرَمَّزة باستخدام لُغة «XML».

٦- المُدَوَّنات اللُّغَوِيَّة وآليَّة فهرسة النُّصُوص

تتطلَّبُ مُعالجَةُ المَدَوَّنات اللُّغَوِيَّة فهرسَةَ نُصُوصِها آليًّا لإتاحة فُرصة التَّعامُل مع الكلمات المفرَدة أو التَّعبيرات المرَكَّبة - على تعدُّد أشكالها وأنماطها؛ وهذا ما تقومُ به المفَهرِسات الآليَّة (Automatic Indexers)، حيثُ تُعنى بتحلِيل نُصُوص المدَوَّنات اللُّغَوِيَّة إلى فِقراتٍ وجُمَلٍ، ثُمَّ تحليل الجُمَلِ إلى كلماتٍ وتعبيرات، ثُمَّ تحليل الكَلِمات إلى فُرُوعٍ (Lemmas / Lemmata) تُمَثِّلُ الوحدات المعجَمِيَّة للنُّصُوص.

وتَخضَعُ عَمَلِيَّة التَّفريع (Lemmatization) لِطَبيعةِ اللُّغةِ المفَهرَسة، فتَزيدُ قُدرةُ المفَهرِسات الآلِيَّة على تَفرِيع الكلِمات عندَ فَهرَسة نُصُوص اللُّغات الإلصاقِيَّة (كالألمانِيَّة والإنجلِيزيَّة والفَرَنسِيَّة)، وتَقِلُّ عند فَهرَسة نُصُوص اللُّغات الاشتِقاقِيَّة (كالعَرَبِيَّة والعِبرِيَّة والحَبَشِيَّة)، وهو أمرٌ تفرضُه الطَّبيعةُ المعجَمِيَّةُ للُّغات الاشتقاقيَّة نظرًا لتعدُّد الوحدات المعجميَّة (Lexemes) الَّتي تتكوَّنُ عنها المفرَداتُ داخلَ الحُقُول المعجَمِيَّة (Entries)؛ لهذا فكثيرًا ما تُستَخدَمُ أدواتٌ لُغَوِيَّةٌ مُساعِدةٌ للمفهرِساتِ الآلِيَّةِ عند تحلِيل النُّصُوص العَرَبِيَّة كالمحَلِّلات الصَّرفِيَّة والتَّركيبِيَّة.



الشَّكل ٢-٥: آلِيَّة التَّفريع Lemmatiser في الإنجليزيَّة والعَرَبِيَّة / Concordance ٣,٣ [1].

1- http://www.concordancesoftware.co.uk.

وتُمثِّلُ آليَّةُ فهرسة النُّصُوص إحدى الرَّكائِزِ الأساسيَّة الَّتي يُستَفادُ منها في مُعالَجة المدَوَّنات اللُّغَويَّة؛ حيثُ تُساعدُ في إدارةِ النُّصُوصِ وحصر تَرَدُّداتِ الكلمات وإعادة تشكيلِها في قواعِد بياناتٍ مُنتَظِمَةٍ؛ كما تُساعِدُ في تعيين الكلمات الفريدة / غير المكَرَّرة (Unique Words) في النُّصُوص، الأمرُ يُوجِّهُ إلى إمكانيَّة التَّعامُل مع مجموعةٍ من الكلمات الَّتي يقلُّ عَددُها كثيرًا عن العدد الكامل لكلمات المدَوَّنة.

ولبيان المعلومات الَّتي يُمكن أن تُوَفِّرَها الآليَّةُ في هذا الصَّدَد، تمَّ إخضاعُ مُدَوَّنةٍ لُغَويَّةٍ مجموعةٍ من نُصُوص الأدب والصِّحافة في العربيَّة المعاصِرة لآليَّة فهرسة النُّصُوص؛ وأفادَت الآليَّةُ أنَّ عددَ الكلمات الفريدة / الَّتي لم تتكرَّر في المدَوَّنة قد بلغَ ١٤٠٥٤٢ كلمة من مجموع كلمات المدَوَّنة البالغ عددها ١٢٢١٩٢٠ كلمة، بنسبة ١١,٥٪؛ كما أفادَت الآليَّةُ أنَّ ما يزيد على نِصف الكلمات الفريدة لم يرِد إلاَّ مرَّةً واحدةً فحسب، وأنَّ أقلَّ من واحدٍ بالمئة من هذه الكلمات الفريدة قد ورَدَ أكثرَ من مئة مرَّة (من ١٠١ إلى ١٠٠٠ بنسبة ٠,٩٪، وأكثر من ١٠٠٠ بنسبة ٠,٠٧٪)، وغير ذلك من النَّتائِج المبيَّنة في (الجدول ٢-٥)؛ ومع أنَّ هذه النَّتائِج ليسَت ثابتةً على صُورَتِها - إذ تتغيَّرُ من مُدَوَّنةٍ لأخرى - إلاَّ أنَّها تُعطِي صُورةً قريبةً للشَّكل الَّذي تكونُ عليه نتائج الفهرسة الآليَّة لنُصُوص المدَوَّنات اللُّغَوِيَّة العربيَّة عُمومًا.

التَّكرار	عدد الكلمات	النِّسبة إلى غير المكَرَّر
١	٧٥٢٧١	٥٣,٥٪
٢ إلى ١٠	٥٢١٣٢	٣٧٪
١١ إلى ١٠٠	١١٧٧٠	٨,٣٪
١٠١ إلى ١٠٠٠	١٢٦٧	٠,٩٪
أكثر من ١٠٠٠	١٠٢	٠,٠٧٪
عدد كلمات المدَوَّنة = ١٢٢١٩٢٠		
عدد الكلمات الفريدة «Unique Words» = ١٤٠٥٤٢ (١١,٥٪)		

الجدول ٢-٦: نتائج الفهرسة الآليَّة لمدَوَّنةٍ لُغَوِيَّةٍ مجموعةٍ من العربيَّة المعاصِرة.

وتجدُرُ الإشارةُ إلى وُجودِ آلِيَّتَين لفهرسة نُصُوص المدَوَّنات اللُّغَوِيَّة؛ تُعرَف الأولى

بالفَهرَسة الألفبائِيَّة؛ وتُتيحُ هذه الآليَّةُ ترتيبَ المفردات ألفبائيًّا دونَ النَّظر إلى السَّوابق واللَّواحق ودونَ مُراعاة الطَّبيعة الاشتقاقيَّة لبعض اللُّغات، لاسيَّما اللُّغة العربيَّة.

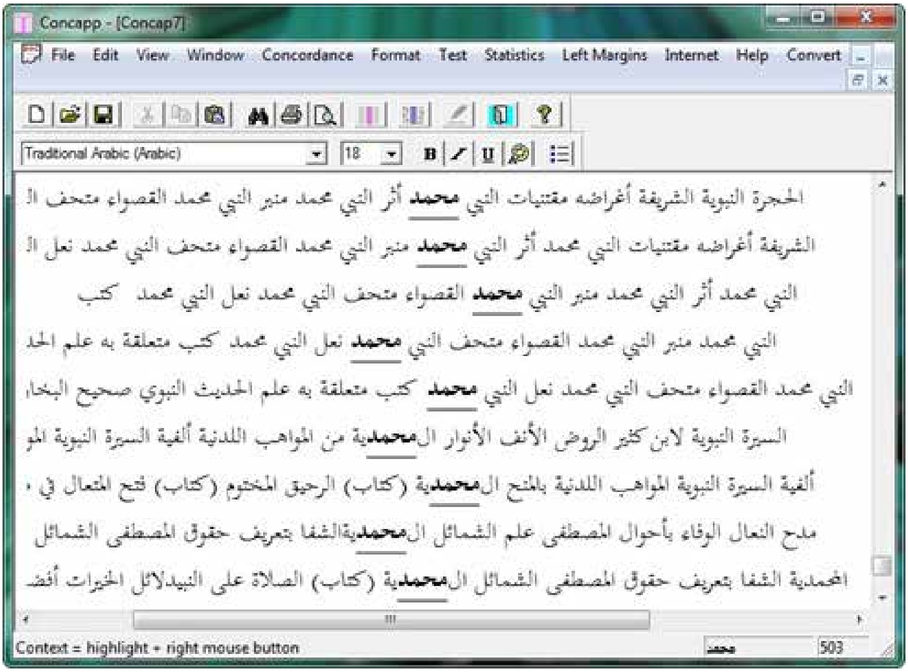


الشَّكل ٢-٦: نموذَجٌ للمفهرِس الآليّ الألفبائيّ aConCorde-٠,٤,٢ [1].

وتُعرَفُ الآليَّةُ الأخرى بالفَهرَسة الجذعِيَّة؛ والجذعُ Stem جُزءٌ من الكلمة، يأتي مُشتَقًّا أو جامِدًا، ويَنتُج عَن اتِّحاد المورفِيمات المكَوِّنة لِبنيَة الكَلِمة الأساسِيَّة (ومِن أمثِلَتِه: الجِذْع «عاودَ» الَّذي تَكَوَّنَ عَنه التَّركِيب «فَعاوَدتْهُم»، «والجِذع «مكتب» الَّذي تَكَوَّنَت عَنه صِيغَة الجَمع «المكتَبات».

وتُتِيحُ هذه الآليَّةُ إمكانيَّةَ البحث عن جذوع الكلمات، بعدَ تجريدِها من السَّوابق واللَّواحِقِ، على نحو ما يُبيِّنُ الشَّكل الآتي:

1- http://www.andy-roberts.net/software/aConCorde/..



الشَّكل ٢-٧: نموذَجٌ للمفَهرِس الآلِيّ الجذعيّ ٠, ٥ Concapp [1].

٧- مجالات الإفادة من المَدَوَّنات اللُّغَوِيَّة

يُمكن الإفادةُ من المَدَونات اللُّغَوِيَّة في الدَّرس اللُّغَوِيّ / علم اللُّغة (Linguistics) وصناعة المعجمات (Lexicography) وتعليم اللُّغات [للنَّاطِقِينَ بها، والنَّاطِقِـينَ بغَيرِها] بالإضافة إلى الدَّور الرَّئِيس للمدَوَّنات اللُّغَوِيَّة في مُعالَجَة اللُّغات الطَّبِيعِيَّة NLP.

٧, ١- استخدام المَدَوَّنات اللُّغَوِيَّة في الدَّرس اللُّغَوِيّ

يرتبط استخدامُ المَدَوَّنات اللُّغَوِيَّة في دراسة اللُّغة بمناهج البحث في عُلُومها. ويَتَعَيَّنُ علينا - قبل الشُّروع في بناء المَدَوَّنةِ - أن نُحَدِّدَ الهدفَ المنشودَ وطبيعةَ الدِّراسة، إذ على أساسِها نستطيعُ أن نُحَدِّدَ شكلَ المَدَوَّنة اللُّغَوِيَّة ومواصفاتها، بما يتناسبُ مع العلم الَّذي ندرسه، وبما يتناسبُ مع المنهج البحثيّ الَّذي نلتزمه في دراستِنا. وسنعرضُ بإيجازٍ - فيما يلي - لبعض جوانب الإفادة من المَدَوَّنات اللُّغَوِيَّة في عُلُوم اللُّغة.

1- http://wmtang.org/200820/11//concapp-5/.

- الأصوات (Phonetics) : تُستَخدَمُ المدَوّناتُ اللُّغَوِيَّةُ المنطوقة في دراسة جوانب التَّبايُن بين التَّنويعات الصَّوتِيَّة «الألوفونات» (Allophones) النَّاتجة عن تغيُّر أشكال الوحدات «الفونيمات» (Phonemes)، ودراسة الظَّواهر الصَّوتِيَّة لهذه الوحدات في سياق الكلام، مثل «النَّبر» (Stress) و «التَّنغيم» (Intonation).
- الصَّرف (Morphology) : تُستَخدَمُ المدَوّناتُ اللُّغَوِيَّةُ النَّصِّيَّة في دراسة الصِّيَغ الصَّرفِيَّة في اللُّغة عبرَ مُستوياتِها، ودراسة أشكال التَّرابُط بين الوحدات الصَّرفِيَّة المجَرَّدة «المورفيمات» (Morphemes) المكَوِّنة لأقسام الكلام، ودراسة أساليب التَّوليد الصَّرفيّ للكلمات المشتَقَّة من الأفعال والمصادر.
- التَّركيب (Syntax) : تُستَخدَمُ المدَوّناتُ اللُّغَوِيَّةُ النَّصِّيَّة أو المنطوقةُ في التَّحليل التَّركيبيّ للجُملة؛ ومن جوانب الإفادة منها: حصر وإحصاء الأنماط التَّركيبيَّة للجُملة في لُغةٍ مُعَيَّنة، والتَّقعيدُ النَّحويّ للُّغات اعتمادًا على الواقع اللُّغَوِيّ المكتوب والمنطوق، ودراسةُ الظَّواهر التَّركيبيَّة الَّتي يُعتَمَدُ عليها في وصف اللُّغة أو استخدامها على النَّحو الَّذي توجدُ عليه بين أهلِها.
- الدِّلالة (Semantics) : تُستَخدَمُ المدَوَّناتُ اللُّغَوِيَّةُ في التَّحليل الدِّلاليّ للنُّصُوص وفي بناء قواعد بيانات الأُنْطُولُوجيات وشبكات الكلمات والشَّبكات الدِّلاليَّة والرَّبط بين حُقُولِها من خلال العلاقات الدِّلاليَّة بين المفرَدات؛ كما تُستَخدَمُ في تَعيين أوجُه الالتباس الدِّلاليّ في النُّصُوص سعيًا إلى إيجاد الوسائل المناسبة الَّتي تُساعدُ في حلِّ مُشكلات هذا الالتباس.
- المعجمِيَّة (Lexicology) : تقومُ المدَوَّناتُ اللُّغَوِيَّةُ بدورٍ فعَّالٍ في تقييم نظريَّات التَّحليل المعجَمِيّ ووضع المناهج الَّتي تُساعِدُ المعجَمِيِّين في بناء المعجَمات والأطالس اللُّغَوِيَّة بما يتناسَبُ مع طبيعة اللُّغة ونظامِها المعجَمِيّ.

وتختلفُ طبيعةُ المدَوَّنات اللُّغَوِيَّة الَّتي تُستَخدَمُ في أيٍّ من هذه العُلُوم الخمسة باختلاف منهج البحث اللُّغَوِيّ الَّذي يسلُكُه الباحثون في دراساتِهم.

- فالدِّراسةُ الوصفِيَّة تتطلَّبُ مُدَوَّنةً لُغَوِيَّةً مُستَمَدَّةً من واقع اللُّغة المكتوبة أو المنطوقة، بصرف النَّظَر عن معايير الصَّواب والخطأ في الاستعمال اللُّغَوِيّ

للنُّصُوص، إذ يُعَوَّلُ على اللُّغة في واقِعِها الحادث؛ وبعبارةٍ أخرى.. تعتمدُ الدِّراسةُ الوصفِيَّةُ للُّغة على ما هو موجود، لا ما ينبغي أن يكونَ موجودا.

- والدِّراسة المِعياريَّة تتطلَّبُ مُدَوَّنةً لُغَوِيَّةً من مُستوى اللُّغة الفصِيحة، لأنَّ هذا النَّوعَ من الدِّراسات لا يسعى إلى وضع القواعد اللُّغَوِيَّة أو التَّنظير لها، وإنَّما يسعى إلى التَّحقُّق من فرضِيَّاتٍ ونظريَّاتٍ لُغَوِيَّةٍ موجودةٍ على أرض الواقِع.
- والدِّراسة التَّاريخيَّة تتطلَّبُ مُدَوَّنةً لُغَوِيَّةً مُمتَدَّةً عبرَ الحُدُود الزَّمانيَّة للُّغة، لأنَّها تهدف إلى دِراسةِ أوجُه التَّبايُن بين مراحل اللُّغة، وما تنفرِدُ بِه كُلُّ مرحلة عن المراحِل الأخرى، وما يطرأ عليها من تَحَوُّلٍ وتَغَيُّر.
- والدِّراسة التَّقابُلِيَّة تتطلَّبُ مُدَوَّنةً لُغَوِيَّةً مُتَعَدِّدَة اللُّغات، لأنَّها تهدفُ إلى المقابَلة بين خصائِص اللُّغة في نظامَين لُغَوِيَّين بهدف تعلُّم اللُّغات، بحيثُ يُمَثِّلُ أحدُ النِّظامَين لُغةَ المتَعلِّم الَّتي يُتقِنُها ويُمَثِّل النِّظامُ الآخرُ اللُّغةَ الَّتي يسعى إلى تَعلُّمِها.

٧,٢- استخدام المدَوَّنات اللُّغَوِيَّة في صناعة المعجمات

اعتمَدَت المعجماتُ القديمةُ في تسجِيل المفرَدات ومعانيها على اللُّغة الَّتي تُستَخدَمُ على ألسنة الجماعة اللُّغَوِيَّة؛ وأَوْلى المعجَمِيُّونَ المتأخِّرُونَ عنايةً بتهذيب المعجمات القديمة وإعادة صياغتِها لتتناسَبَ مع اللُّغة المعاصِرة ومُستَجَدَّاتها الَّتي تفرضُها عواملُ التَّطَوُّر اللُّغَوِيّ المتأثِّرة بالزَّمان والمكان والأحداث.

لكنَّ واقِعَ الصِّناعة المعجَمِيَّة في كثيرٍ من اللُّغات يُؤَكِّدُ أنَّ المعجَمِيِّين لم يُوَفَّقُوا إلى هذا الهدفِ، لاعتمادِهِم على المعجمات القديمة والدِّراسات المعجَمِيَّة فحسب، دونَ النَّظَر إلى واقع اللُّغة الَّذي لا يُمكنُ التَّعبيرُ عنه إلاَّ من خلال النُّصُوص المستَمَدَّة من هذا الواقع، الأمرُ الَّذي أدَّى إلى عدم التَّمييز بين المهمَل والمستَعمَل من مُفرَدات اللُّغة ومعانيها. ونستطِيعُ أن نُمَثِّلَ على ذلك من اللُّغة العربيَّة بالمعجم الوسيط (وهو مُعجمٌ عربيٌّ مُعاصِرٌ، أصدَرَهُ مجمعُ اللُّغة العربيَّة بالقاهرة في طبعتِه الأولى عامَ ١٩٦٠م، وفي طبعته الثَّالثة عامَ ٢٠٠٣م) حيثُ يُعنى بإيراد معاني العديد من الحُقُول المهمَلة، على شاكلة (بجبجَ، وبَحشَلَ، وجَعبَبَ) ولا يُعنى بكلمتَيْ (حاسِب، وحاسُوب) الشَّائِعَتَين.

وهُنا تظهرُ فائدةُ استخدام المدَوَّنات اللُّغَوِيَّة في صناعة المعجَم، لأنَّ المعجَمِيَّ - حينَئِذٍ - سيجمعُ المادَّةَ الَّتي يعتمدُ عليها في صناعة المعجَم من اللُّغة المستعمَلة والمتَدَاولة بينَ أهلِها، لا من اللُّغة المهجورة في ثنايا المعجمات القديمة؛ كما سيكونُ قادرًا على التَّمييز بينَ المستَعمَل والنَّادرِ والمهمَلِ من المفرَدات والمعاني. ومع تحقُّق هذه الفائدة في اللُّغة الإنجليزِيَّة وفي العديد من اللُّغات الجِرمانِيَّة واللاَّتِينِيَّة، إلاَّ أنَّها لم تتحقَّق بعدُ في اللُّغة العربيَّة الَّتي تُعاني نقصًا كبيرًا في مُعجماتِها المعاصِرة [المكتوبة أو المنطوقة] ومُعجمات اللَّهجات والمعجَمات الاصطلاحيَّة والتَّعليميَّة والتَّاريخِيَّة.

٧,٣- استخدام المدَوَّنات اللُّغَوِيَّة في تَعليم اللُّغات

مَهَّدَت تجربةُ عالِمِ النَّفس الأمريكيّ «إدوارد ثورنديك» - سالفة الذِّكر - الطَّريقَ إلى استخدام المدَوَّنات اللُّغَوِيَّة في تعليم اللُّغات. ونتجَ عن هذه التَّجربة ظُهُورُ ما يُعرَفُ بـ «قوائم الكلمات الشَّائِعة» (Lists of common words) في العديد من اللُّغات، بما في ذلكَ العربيَّة الَّتي عَرَفَت هذه القوائم من خلال أعمال اللُّغَوِيِّين والتَّربويِّين المعنِيِّين بتعليم اللُّغة العربيَّة. وكانَ من هذه الأعمال -على سبيل المثالِ لا الحَصر-: «قائمة المفردات الشَّائعة في العَربيَّة الحديثة» (A list of Modern Arabic words) الَّتي أعَدَّها «إوينج بيلي» (Ewing Macready Bailey) في عام ١٩٤٨، والدِّراسة الموسُومة بـ «المفردات الشَّائعة في اللُّغة العربيَّة» لداود عبده / في ١٩٧٩، والدِّراسة الموسُومة بـ «قائمة المفردات الشَّائعة الاستخدام في البلاد العربيَّة» لرُشدي طعيمة / في ١٩٨٢.

ومع أنَّ المدَوَّنات اللُّغَوِيَّة الَّتي يُعتَمَدُ عليها في إعداد مثلِ هذه القوائم لا تُعَبِّرُ بالضَّرورة عن واقعِ اللُّغة-إذ يتمُّ اختيارُ نُصُوصِها عشوائِيًّا في كثيرٍ من الأحيان، إلاَّ أنَّ الفكرةَ ذاتها تُمَثِّلُ وسيلةً عمليَّةً يُمكنُ تطويرُها منهجيًّا، حيثُ تُوَجِّهُ مُتَعَلِّمي اللُّغة إلى التَّعرُّف أوَّلاً على المفرَدات الأكثر شُيُوعًا في نطاق المستوى اللُّغَوِيّ الَّذي تُعنى به دراسَتُهم، ثُمَّ الانتقال إلى المفرَدات الأقلِّ شُيُوعًا، ثُمَّ الَّتي تليها، وهكذا.

ولا تقتصرُ العمليَّاتُ الإحصائِيَّةُ على المدَوَّنات المنجَزة لأغراضٍ تعليميَّةٍ على الكلمات؛ وإنَّما تمتدُّ لتشملَ الأنماطَ البِنَوِيَّة والتَّركيبيَّة للُّغة، والَّتي تُساعدُ على تعلُّم القوعد النَّحويَّة. ففي اللُّغة العربيَّة -مثلاً - يُمكنُ الاعتمادُ على المدَوَّنات اللُّغَوِيَّة الممَثِّلة لواقع اللُّغة [على مُستوى البِنية] في حصر الأوزان الصَّرفيَّة الشَّائعة للأفعال

والمشتَقَّات والمصادر بأنواعِها؛ كما يُمكنُ الاعتمادُ عليها في حصر الصِّيَغ الشَّائعة لكُلِّ نوعٍ من المشتَقَّات على حِدَة. أمَّا [على مُستوى التَّركيب] فيُمكن الإفادةُ من المدَوَّنات اللُّغَوِيَّة في التَّعرُّف على المواقع الإعرابيَّة الأكثر تردُّدًا وشُيُوعًا؛ وكذلكَ في حَصر أنماط الجُملة العربيَّة، الأمر الَّذي يُساعدُ على معرفة أكثر الأنماط التَّركيبيَّة شُيُوعًا، واستنباط خصائِص الجُملة العربيَّة من حيثُ مُتَوَسِّطات أطوالِها وتتابُعات أقسامِها.

ومن ناحيةٍ أخرى، تُؤَدِّي المدَوَّناتُ اللُّغَوِيَّةُ دورًا كبيرًا في تطوير المناهج التَّعليميَّة للُّغات في مراحل التَّعليم المختلفة، إذ من خلالِها يُمكنُ توجيهُ الطُّلاَّب إلى واقع اللُّغة الملموس، بعيدًا عن التَّعقيدات الَّتي قد لا يُفيدونَ منها في واقِعِهم أو مُستقبَلِهم، فنجنِّبهم بذلكَ الحُوشِيَّ والغَريبَ والمهجور. ولَعَلَّنا نستشعرُ أهمِّيَّةَ ذلكَ بالنَّظَر إلى مُعاناة طُلاَّب المراحل الأساسيَّة في فهم مناهج اللُّغة العربيَّة، لاسيَّما قواعد النَّحو العربيّ. ذلكَ أنَّ الطَّالبَ يكونُ مُلزَمًا بدراسة القاعدة النَّحْويَّة الَّتي يحويها المنهجُ التَّعليميّ، بصرف النَّظَر عن الاستخدام الفِعليّ لها. ونرى أنَّ المادَّةُ العلميَّةَ المقَدَّمةَ - في أحيانٍ كثيرةٍ - تبدو بعيدةً عن اللُّغة الَّتي يُمارسُها الطُّلابُ قراءةً أو استماعًا؛ ناهيكَ عن ضَعف مُستوى المعلِّمينَ نتيجةَ القُصُور في تأهيلِهم، ما يُؤدِّي إلى عدم قُدرَتِهم على التَّواصُل مع طُلاَّبِهم.

ويُمكنُ التَّمثيلُ في هذا الصَّدَد بما نراهُ مُقَرَّرًا على الطُّلابِ في دراستِهم للنَّحو العربيّ. من ذلكَ أنَّ بعضَ المناهجِ التَّعليميَّة تُلزِمُ الطُّلاَّبَ بدراسة باب (كانَ وأخواتها) بكُلِّ ما يحويه من قواعدَ أساسيَّة. ويُوَجَّهُ الطَّالبُ في دراسَتِهِ لهذا الباب إلى الأفعال (ماانفكَّ، مافتِئَ، مابرِحَ). ومع أنَّ هذه الأفعالَ ليست شائعةً شُيُوعَ غيرِها، إلاَّ أنَّ الطَّالبَ يدرسُها لأنَّ القاعدَةَ تُحَتِّمُ عليه دراسَتَها، وإن لم يسمعها من قبل أو يستخدمها؛ ورُبَّما لن يستخدِمَها - كذلكَ - في مُستَقبَلِه.

ومع أهمِّيَّة أن يتعرَّفَ الطَّالبُ على جمال لُغَتِه وعُذُوبَتِها، إلاَّ أنَّ تقديمَ بعضِ القواعد أو الأساليب النَّادرة في مرحلةٍ تعليميَّةٍ أوَّليَّةٍ سوفَ يُؤَثِّرُ بالضَّرورة على مُستوى تحصِيلِه، وسيُؤَدِّي حتمًا إلى إهمال قواعدَ وأساليب أكثرَ شُيُوعًا واستخدامًا ممَّا وُجِّهَ إليه. وحينئذٍ لا يكونُ مُستغربًا أن يقرأَ فلا يفهم، أو ينطقَ فلا يُحسِن.

ولو أنَّ المدَوَّنات اللُّغَوِيَّة وُظِّفَت لحصر الشَّائع [المستَخدم فِعليًّا] مِن المفرَدات والتَّراكيب والأساليب، ثُمَّ توجيه الطُّلاَّب إليه أوَّلاً، لأمكنَ الارتقاءُ بمُستواهم

التَّعليميّ إلى درجةٍ كبيرة، إذ يُوَجَّهُونَ حِينَئِذٍ إلى قواعدَ مُستَمَدَّةٍ من واقِعِ لُغَتِهِم الَّذي يُعاصِرُونَه، سواءٌ أكانَت هذه اللُّغة مكتوبة أم منطوقة.

٧,٤ – استخدام المدَوَّنات اللُّغَوِيَّة في مُعالَجَة اللُّغات الطَّبيعيَّة

ثَمَّةَ العديدُ من تطبيقات مُعالَجَة اللُّغات الطَّبيعيَّة الَّتي تعتمِدُ على المدَوَّنات اللُّغَوِيَّة في مراحل إعدادها وتقييمها؛ ويُحَدِّدُ نوعُ المدَوَّنة اللُّغَوِيَّة المستَخدَمة وخصائِصُها وفقًا لطبيعة التَّطبيق والهدف المنشود منه. وسنُحاوِلُ الوُقُوفَ على دور المدَوَّنات اللُّغَوِيَّة في مُعالَجَة اللُّغات الطَّبيعيَّة من خلال بعض التَّطبيقات على النَّحو التَّالي:

- التَّرجمة الآليَّة: وأداتُها الأساسيَّةُ مُدَوَّنةٌ لُغَوِيَّةٌ مُتَوازية [أو مُتَقارِبة] بين اللُّغات الَّتي يُعالِجُها نظام التَّرجمة المنشود. ويُستَفادُ منها في تعيين قواعد النَّحو وأنماط التَّراكيب الَّتي تُهَيِّئُ النِّظامَ للتَّرجمة النَّصِّيَّة، كما يُستَفادُ منها في إثراء قاعدة بيانات النِّظام بمُفرَدات اللُّغة المصدر وما يُقابلُها من مُفرَدات اللُّغة الهَدَف.
- آليَّة التَّدقِيق الإملائيّ: ويتَطَلَّبُ إعدادُها مُدَوَّنَتَين لُغَوِيَّتَين، تُعنى الأولى بمرحلة التَّدريب؛ وينبغي أن تكونَ خاليةً من الأخطاءِ الإملائِيَّة ليُستَفادُ منها في إثراء مُعجم الآليَّة وقواعد بيانات النِّظام البرمَجِيّ؛ وتُعنى الأخرى بمرحلة التَّقييم، وتُجمَعُ مادَّتُها من نُصُوصٍ عشوائِيَّةٍ، ليُستَفادَ منها في اختبار الآليَّة وتَعيين مُعَدَّل الخطأ في نتائِجِها.
- آليَّة تشكيل النُّصُوص: وهي تطبيقٌ خاصٌّ باللُّغة العربيَّة الَّتي تنفرِدُ بظاهرة الإعراب وتتمتَّعُ بنظام كتابيٍّ يُمَيِّزُها عن غيرِها من اللُّغات. يتَطَلَّبُ إعدادُ هذه الآليَّة مُدَوَّنةً لُغَوِيَّةً مشكولةً كُلِّيًّا، بحيثُ يُمكن الإفادةُ منها في تحديد الطُّرُق الإحصائِيَّة الَّتي ستعتمدُ عليها خوارزمات التَّشكيل.
- آليَّة فكّ الالتِباس الدِّلاليّ: ويتَطَلَّبُ إعدادُها مُدَوَّنةً لُغَوِيَّة غَنيَّةً بالمتَصاحبات اللُّغَوِيَّة والتَّعبيرات الاصطلاحيَّة والكلمات الَّتي تحمل دلالاتٍ مُتَعَدِّدة. ويُستَفادُ من هذه المدَوَّنة في إثراء قواعد بيانات النِّظام المنشود بالمفردات ودلالاتِها، وتدريب خوارزمات فكّ الالتباس باستِدعاءِ معاني الكلمات ذات الدِّلالات المتَعَدِّدةِ من الكَلِمات المصاحِبَة لها في سياق النُّصُوص.

٨- أفكارٌ بحثيَّة لأطرُوحاتٍ علميَّةٍ مُستَقبليَّة

نَظَرًا لنُدرة الدِّراسات الَّتي كُتِبَت بالعَرَبِيَّة عن المدَوَّنات اللُّغَوِيَّة وحداثة منهج دراستها على العربيَّة ولهجاتها، سنُحاوِلُ -فيما يلي- أن نعرِضَ لبعضِ الأفكار البحثِيَّة الَّتي قد تَصلُحُ لإنجاز أطرُوحاتٍ علميَّةٍ للباحِثين العَرَب.

٨,١- موضُوع الفِكرة الأولى:

التَّطوُّر اللُّغَوِيّ في لُغة الصِّحافة المِصرِيَّة المعاصِرة

«دراسة إحصائيَّة في ضوء مُدَوَّنة لُغَوِيَّة»

- **مادَّة الدِّراسة:**

مُدَوَّنةٌ لُغَوِيَّةٌ مكتُوبة (نَصِّيَّة) مُستَمَدَّةٌ من نُصُوص الصِّحافة المِصرِيَّة المعاصِرة.

- **الأسئِلة البَحثِيَّة:**
  - ما الخُطُواتُ المنهجيَّةُ لبناء مُدَوَّنةٍ لُغَوِيَّةٍ للصِّحافة المِصرِيَّة المعاصِرة؟
  - ما أساليبُ التَّحليل الإحصائِيّ لنُصُوص المدَوَّنة اللُّغَوِيَّة موضُوع الدِّراسة؟
- **منهج الدِّراسة، ومجال البحث:**

تقومُ الدِّراسةُ المقتَرَحةُ على المنهجين: الوصفيّ والتَّارِيخيّ، ويَتَنَوَّعُ مجالُ البحث بين الإحصاء اللُّغَوِيّ ولسانِيَّات المدَوَّنة.

- **المراجع الأوَّلِيَّة المقتَرَحة:**

عبد العزيز (محمد حسن): لغة الصحافة المعاصرة، دار الفكر العربي، القاهرة، ط١، ٢٠٠٢م.
Patten, M. L. (2007). Understanding research methods: an overview of the essentials. Pyrczak.

٨,٢- موضُوع الفِكرة الثَّانية:

الأطلس اللُّغَوِيّ للعربيَّة الدَّارِجة في مِصر

«دراسة وصفيَّة في ضوء مُدَوَّنة لُغوِيَّة منطوقة»

- **مادَّة الدِّراسة:**

مُدَوَّنةٌ لُغَوِيَّةٌ منطوقة مُستَمَدَّةٌ من نُصُوص اللُّغة العربيَّة في مصر، أو في إحدى اللَّهجات المِصريَّة المعاصِرة.

- **الأسئِلة البَحثِيَّة:**
  - ما المعايير الَّتي ينبغي توافُرُها في الأطلس اللُّغَوِيّ المنشود؟
  - ما الخُطُواتُ المنهجيَّةُ لبناء مُدَوَّنةٍ لُغَوِيَّةٍ منطوقةٍ للعربيَّة الدَّارِجة في مِصر؟
  - كيفَ يُمكِن الإفادةُ من الأطلس اللُّغَوِيّ للعربيَّة الدَّارِجة في مِصر في دراسة الظَّواهر اللُّغَوِيَّة للعامِّيَّة المِصريَّة المعاصِرة؟
  - كيفَ يُمكن توظيفُ المدَوَّنة اللُّغَوِيَّة المنطوقة [أداة الدِّراسة] في بناء أطلس لُغَوِيّ للعربيَّة الدَّارِجة في مِصر؟

- **منهج الدِّراسة، ومجال البحث:**

تقومُ الدِّراسةُ المقتَرَحةُ على المنهج الوصفيّ، ومجالُ البحث لسانِيَّات المدَوَّنة.

- **المراجع الأوَّلِيَّة المقتَرَحة:**

عساكر (خَليل): الأطلس اللُّغَوِيّ، مجلَّة مجمع اللُّغة العربيَّة بالقاهرة، الجُزء السَّابع، ص ٢٨١ - ٣٨٣، ١٩٤٩م.
Bergsträßer, G. (1995) Sprachatlas von Syrien und Palästina. Leipzig, J.C. Hinrichs.

٨,٣- موضُوع الفِكرة الثَّالثة[١]:

المعجَم التَّكراريّ للُّغة العربيَّة المعاصِرة
«المنهج والنَّموذج في ضوء مُدَوَّنة لُغَوِيَّة»

• **مادَّة الدِّراسة:**

مُدَوَّنةٌ لُغَوِيَّةٌ مكتُوبة (نَصِّيَّة) مُستَمَدَّةٌ من نُصُوص اللُّغة العربيَّة المعاصِرة، وتتنوَّعُ مادَّتُها لتشمَلَ: الآداب العربيَّة، ولُغة الصِّحافة، والمعارِف العامَّة.

• **الأسئِلة البَحثِيَّة:**

• ما المقصُودُ بالمعجَمات التَّكرارِيَّة؟ وما مكانتُها في اللُّغة العربيَّة؟

• كيفَ يُستَفادُ من المعجم المنشود في تعليم العربيَّة لغير النَّاطِقين بها؟

• **منهج الدِّراسة، ومجال البحث:**

تقومُ الدِّراسةُ المقتَرَحةُ على المنهج الوصفيّ، ويَتَنَوَّعُ مجالُ البحث بين الإحصاء اللُّغَوِيّ ولسانِيَّات المدَوَّنة وعلم اللُّغة الحاسُوبيّ.

• **المراجع الأوَّلِيَّة المقتَرَحة:**

السَّعيد (المعتزّ بالله): نحوَ مُعجم للُّغة العربيَّة للنَّاطِقينَ بغيرها «مُعالجة حاسُوبيَّة إحصائيَّة»، مجلَّة «التَّواصُل اللِّسانيّ» –المجلَّة الدَّوليَّة لهندسة اللُّغة العربيَّة واللِّسانيَّات العامَّة، LINGUISTICA COMMUNICATIO (International journal of Arabic Language Engineering & General Linguistics)، فاس، المغرب، مج ١٩، ٢٠١٨.

١- أمكنَ تنفيذُ هذه الفكرة في رسالة ماجستير بعُنوان (المعجم التَّكراريّ للُّغة العربيَّـــة: مُعالجة لُغَوِيَّة حاسُوبيَّة)، للباحث مُحمَّد مجدي، بإشراف المؤلِّف.

٨,٤- موضُوع الفِكرة الرَّابعة:

تقييم أدوات التَّحليل التَّركيبيّ في اللُّغة العربيَّة

«دِراسةٌ لُغَوِيَّة حاسُوبيَّة في ضوء مُدَوَّنة لُغَوِيَّة مُعَنونة»

- **مادَّة الدِّراسة:**

مُدَوَّنةٌ لُغَوِيَّةٌ مكتُوبة (نَصِّيَّة) ومُعَنونة تركيبيًّا؛ مُستَمَدَّةٌ من نُصُوص اللُّغة العربيَّة المعاصِرة، وتتنوَّعُ مادَّتُها بين الآداب العربيَّة ولُغة الصِّحافة.

- **الأسئِلة البَحثِيَّة:**
  - كيفَ نبني مُدَوَّنةٍ لُغَوِيَّةٍ لتقييم المحلِّلات التَّركيبيَّة العربيَّة؟
  - إلى أيِّ مدًى يُمكن الإفادةُ من آليَّة التَّحليل التَّركيبيّ في العربيَّة؟
  - ما أهمّ الأساليب الإحصائِيَّة الَّتي يُمكن الإفادة منها في مراحل التَّقييم؟

- **منهج الدِّراسة، ومجال البحث:**

تقومُ الدِّراسةُ المقتَرَحةُ على المنهج الوصفيّ، ويَتَنَوَّعُ مجالُ البحث بين الإحصاء اللُّغَوِيّ ولسانِيَّات المدَوَّنة وعلم اللُّغة الحاسُوبيّ.

- **المراجع الأوَّلِيَّة المقتَرَحة:**

حَسَّان (تمَّام): اللُّغة العَرَبيَّة «مَعناها وَمَبناها»، الهَيئة المِصريَّة العامَّة للكِتاب، القاهِرة، ط٢، ١٩٧٩م.
Attia, M. & Rashwan, M., A Large-Scale Arabic POS Tagger Based on a Compact Arabic POS Tags Set, and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words, The Proceedings of the Arabic Language Technologies and Resources Int'l Conference; NEMLAR, Cairo-Egypt http://www.elda.org/nemlar-conf, Sept. 2004.

٨,٥- موضُوع الفِكرة الخامسة:

تقييم نُظُم التَّرجمة الآليَّة الحديثة بين العربيَّة والإنجليزيَّة
«في ضوء مُدَوَّنة لُغَوِيَّة مُتَوازِية»

- **مادَّة الدِّراسة:**

مُدَوَّنةٌ لُغَوِيَّةٌ مكتُوبة (نَصِّيَّة) ومُتَوازِية، مُستَمَدَّةٌ من لُغة الصِّحافة المعاصِرة. تُمَثِّلُ العربيَّةُ فيها (اللُّغَةَ المصدر)، وتُمَثِّلُ الإنجليزِيَّة (اللُّغَةَ الهَدف).

- **الأسئِلة البَحثِيَّة:**
  - ما الخُطوات المنهجيَّة لبناء مُدَوَّنةٍ لُغَوِيَّة مُتوازية للعربيَّة والإنجليزيَّة؟
  - ما أهمّ الأساليب الإحصائِيَّة الَّتي يُمكن الإفادة منها في مراحل التَّقييم؟

- **منهج الدِّراسة، ومجال البحث:**

تقومُ الدِّراسةُ المقتَرَحةُ على المنهجين: الوصفيّ والتَّقابُليّ، ويَتَنَوَّعُ مجالُ البحث بين الإحصاء اللُّغَوِيّ ولسانِيَّات المدَوَّنة وعلم اللُّغة الحاسُوبيّ.

- **المراجع الأوَّلِيَّة المقتَرَحة:**

Joseph Olive (2011): Handbook of Natural Language Processing and Machine Translation. Springer.
Szymon Rutkowski (2012): Machine Translation Evaluation: An Analysis of Two Translations Produced by Google Translate and English Translator XT. Lambert Academic Publishing.
Dehcheshmeh, M. (2007). Specialized Monolingual Corpora in Translation. Translation Journal. Volume 11, No. 2.
Carmen Mill N-Varela, Francesca Bartrina (2013): The Routledge Handbook of Translation Studies. Routledge.

٩- مِن المواقع الإلكترونيَّة التَّعليميَّة والإرشاديَّة

1- http://corpora.wordpress.com/

- موقع تمهيديّ، يُعرِّف بالمدَوَّنات اللُّغَوِيَّة، ويَعرِض مجموعة من الأدوات المستَخدمة في مُعالَجَة نُصُوصِها آليًّا.

2- http://arabicorpus.byu.edu/

- موقع المدَوَّنة اللُّغَوِيَّة العربيَّة، يستمدُّ مادَّته من الصُّحُف العربيَّة، ويُمكن الإفادة منه في أغراضٍ بحثيَّة مُختلفة.

3- http://corpus.byu.edu/

- موقع المدَوَّنات اللُّغَوِيَّة الأنجلوأمريكيَّة، ويَضُمُّ-كذلك- مُدَوَّنَتَين لُغَوِيَّتَين للإسبانيَّة والبُرتغاليَّة.

4- http://faculty.washington.edu/ebender/corpora/corpora.html

- موقع الموقع مجموعةً من روابط المدَوَّنات اللُّغَوِيَّة للُغاتٍ عديدة، ويُعنى بروابط مواقع المشروعات اللُّغَوِيَّة الكُبرى.

5- http://www.uncorpora.org/.

- موقع المدَوَّنات اللُّغَوِيَّة للأمم المتَّحدة، ويضُمُّ مجموعةً من الوثائق الَّتي يُمكن الإفادة منها في أغراضٍ بحثيَّة مُختلفة.

6- http://www.natcorp.ox.ac.uk/.

- موقع المدَوَّنة الوطنيَّة البريطانيَّة، يُمكن البحث فيه عن المفرَدات، وتتنوَّعُ مادَّة المدَوَّنة بين المكتوب والمنطوق.

7- http://www.comp.leeds.ac.uk/eric/latifa/index.htm.

- موقع الباحِثة القَطَرِيَّة لطيفة السّليطي على موقع جامعة ليدز، يَضُمُّ قائِمةً ببعض المدَوَّنات اللُّغَوِيَّة العربيَّة وتعريفًا موجَزًا بها

ببليوجرافيا مرجعيَّة

١. السَّعيد (المعتزّ بالله): مُدَوَّنة مُعجَم عربيّ مُعاصِر: مُعالَجة لُغَوِيَّة حاسوبِيَّة، رسالة ماجستير، جامعة القاهرة، ٢٠٠٨م.

٢. السَّعيد (المعتزّ بالله): مُدَوَّنة مُعجَم تاريخيّ للُّغة العربيَّة: مُعالَجة لُغَوِيَّة حاسوبِيَّة، أطروحة دكتوراه، جامعة القاهرة، ٢٠١١م.

3. Abdel-Fattah, Y. (2018). Arabic Corpus Linguistic. Edinburgh University Press.

4. Adolphs, S.; Carter, R. (2013): Spoken Corpus Linguistics: From Monomodal to Multimodal. Taylor & Francis Group.

5. Aijmer, k.; Altenberg, B. (2014). English Corpus Linguistics. Routledge.

6. Aijmer, K.; Bengt Altenberg, B. (2013): Advances in Corpus-Based Contrastive Linguistics: Studies in Honour of Stig Johansson. John Benjamins Publishing Company.

7. Al-Sulaiti, L. (2004). Designing and Developing a Corpus of Contemporary Arabic. "M.Sc. thesis". Leeds University.

8. Arulmozi, S.; Dash, N. (2018). History, Features, and Typology of Language Corpora. Springer.

9. Baker, P. (2012): Contemporary Corpus Linguistics. Bloomsbury.

10. Baker, P.; Hardie, A.; McEnery, T. (2006). A Glossary of Corpus Linguistics. Edinburgh University Press.

11. Biber, D.; Reppen, R. (2015): The Cambridge Handbook of English Corpus Linguistics. Cambridge University Press.

12. Brezina, V. (2018). Statistics in Corpus Linguistics: A Practical Guide. Cambridge University Press.

13. Collinge, N. E. (2013): Encyclopaedia of Language. Taylor & Francis.

14. Collins, L. (2019). Corpus Linguistics for Online Communication: A Guide for Research. Routledge.
15. Crawford, W.; Csomay, E. (2015): Doing Corpus Linguistics. Taylor & Francis Limited.
16. Dillmann, L.; Arndt-Lappe, S.; Sand, A.; Hoffmann, S. (2018). Corpora and Lexis. BRILL.
17. Ender, A.; Leemann, A.; Wälchli, B. (2012): Methods in Contemporary Linguistics. Walter de Gruyter.
18. Eric, A. (2017). Sociolinguistics and Corpus Linguistics. Magnum Publishing.
19. Farr, F.; Murray, L. (2016): The Routledge Handbook of Language Learning and Technology. Routledge.
20. Flowerdew, L. (2012): Corpora and Language Education. Palgrave Macmillan.
21. Friginal, E. (2017). Studies in Corpus-Based Sociolinguistics. Routledge.
22. Gomez, P. C. (2013): Statistical Methods in Language and Linguistic Research. Isd.
23. Gries, S. T. (2009): Quantitative Corpus Linguistics With R: A Practical Introduction. Taylor & Francis.
24. Gries, S. T.; Wulff, S.; Davies, M. (2010): Corpus Linguistic Applications: Current Studies, New Directions. Rodopi.
25. Handford, M. (2018). Corpus Linguistics for Discourse Analysis: A Guide for Research. Routledge.
26. Hansen-Schirra, S.; Neumann, S.; Steiner, E.; Culo, O.; Hansen, S. (2012): Cross-Linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German. De Gruyter.
27. Heine, B.; Narrog, H. (2009). The Oxford Handbook of Linguistic Analysis. Oxford University Press.

28. Hinkel, E. (2013): Handbook of Research in Second Language Teaching and Learning. Volume I. Routledge.

29. Hoffmann, T.; Trousdale, G. (2013): The Oxford Handbook of Construction Grammar. Oxford University Press.

30. Hunston, S. (2002). Corpora in applied linguistics. Cambridge: Cambridge University Press.

31. Koplenig, A. (2017). Against Statistical Significance Testing in Corpus Linguistics. Stefanie Wulff.

32. Ludeling, A.; Kyto, M. (2009) Corpus Linguistics. Walter de Gruyter.

33. Markus, M. (2012): Middle and Modern English Corpus Linguistics: A Multi-Dimensional Approach. John Benjamins Publishing.

34. Marzo, S.; Heylen, K.; Sutter. G. (2012): Corpus Studies in Contrastive Linguistics. John Benjamins Publishing.

35. McEnery, T. (2014). Arabic Corpus Linguistics. Edinburgh University Press.

36. McEnery, T.; Hardie, A. (2011): Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.

37. McEnery, T.; Meurers, D.; Rebuschat, P. (2017). Experimental, Corpus-based and Computational Approaches to Language Learning: Evidence and Interpretation. Wiley.

38. McEnery, T.; Wilson, A. (2001). Corpus Linguistics “An introduction”. 2nd edition. Edinburgh University Press.

39. Mikhailov, M.; Cooper, R. (2016): Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research. Routledge Corpus Linguistics Guides. Routledge.

40. Millán, C.; Bartrina, F. (2013): The Routledge Handbook of Translation Studies. Routledge.

41. O'Keeffe, A.; McCarthy, M. (2012): The Routledge Handbook of Corpus Linguistics. Routledge.

42. Pahta, P.; Rütten, T.; Nurmi, A. (2017). Challenging the Myth of Monolingual Corpora. BRILL.

43. Posch, C. (2014). Feminist Linguistics and Corpus Linguistics: A database of genderfair language use with non-human referents. GRIN Verlag.

44. Ramamoorthy, L.; Dash, N. (2018). Utility and Application of Language Corpora. Springer Singapore.

45. Rass, T. (2013): Corpus Linguistics - An Introduction to the Field and its Use in Linguistics. GRIN Verlag.

46. Romero-Trillo, J. (2016): Yearbook of Corpus Linguistics and Pragmatics 2016: Global Implications for Society and Education in the Networked Age. Springer International Publishing.

47. Rühlemann, C. (2018). Corpus Linguistics for Pragmatics: A guide for research. Routledge.

48. Schiebert, W. (2011): Corpus Linguistics: Lexicography and Semantics: Introduction to Concordance and Collocations. GRIN Verlag.

49. Schmidt, T.; Wörner, K. (2012): Multilingual Corpora and Multilingual Corpus Analysis. John Benjamins Publishing.

50. Sinclair, J. M. (2004). How to use corpora in language teaching. John Benjamins Publishing Company.

51. Speelman, D. (2018). Mastering Corpus Linguistics Methods: A Practical Introduction with AntConc and R. John Wiley & Sons.

52. Walker, B.; Mcintyre, D. (2019). Corpus Stylistics: Theory and Practice. Edinburgh University Press.

53. Weisser, M. (2015): Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis. John Wiley & Sons.

54. Wray, A.; Bloomer, A. (2013): Projects in Linguistics and Language Studies. Routledge.

الفصل الثَّالث
الشَّبكات الدِّلاليَّة

د. سامِح الأنصاري

١ – التحليل الدلالي للجملة: لمحة تاريخية.

٢ – لغة الشبكات الدلالية الحاسوبية العالمية.

٣– المكونات اللغوية للغة الشبكات الدلالية الحاسوبية العالمية.

٤ – موارد وأدوات لغة الشبكات الدلالية الحاسوبية العالمية.

٥ – تطبيقات المعالجة الآلية للدلالة باستخدام لغة الشبكات الدلالية الحاسوبية العالمية.

٦ – دعوة للمشاركة.

يقول عبد القاهر الجرجاني في كتابه (دلائل الإعجاز في علم المعاني): «الألفاظ المفردة التي هي أوضاع اللغة لم توضع لتعرف معانيها في أنفسها، ولكن لأن يضم بعضها إلى بعض فيعرف فيما بينها فوائد» إنّ هذا النص فيه إشارة إلى أن الدلالة لا تقتصر على الجانب الإفرادي فقط وإنما تتعداه إلى الجانب التركيبي النحوي. فلا تتحقق الفائدة بالنظر في معاني الألفاظ المفردة بل تتحقق بمعرفة المعاني الناتجة عن ضم تلك الكلمات بعضها مع بعض. وإذا كان الفصل الأوَّل من هذا الباب قد تعرض للمعالجة الدلالية لمفردات اللغة فإننا في هذا الفصل نصل لأبعد من هذا، وتحديدًا إلى معالجة الدلالة على مستوى الجملة. ولعل التحليل الدلالي للجملة الطبيعية يُعد من أصعب أنواع التحليلات اللغوية لأن تحليل دلالة عبارة ما، لا بد أن يتم على مستويات متعددة هي معنى مفرداتها (المعجم) ومعنى الأجزاء التركيبية لكلماتها (الصرف) ومعنى العلاقات التركيبية بين تلك الكلمات (النحو) والمعاني الخارجة عن التركيب التي تنتج عن ظروف برجماتِيَّة/ غير لغوية (Pragmatic) ولكنها تؤثر في المعنى اللغوي. هذا التعدد في مستويات التحليل جعل محاولات المعالجة الحاسوبية لدلالة الجملة الطبيعية في أبكر مراحلها.

إنّ هذا الفصل ما هو إلا مقدمة لواحدة من المحاولات الطموحة في المعالجة الآلية للدلالة على مستوى الجملة العربية تحليلاً وتوليدًا باستخدام «لغة الشبكات الدلالية الحاسوبية العالمية» (Universal Networking Language–UNL) متضمنة التعريف بتلك اللغة الحاسوبية وتقديم وصف نظري وعملي لطبيعة عملها مع عرض لأبرز التطبيقات التي يمكن للغة الشبكات الدلالية الحاسوبية العالمية أن تساهم في تطويرها. وننتهي أخيرًا بالتعرض لبعضٍ من النقاط البحثية الهامة من أجل دعم خطة طريق لمعالجة الدلالة في الجملة العربية.

١- التحليل الدلالي للجملة: لمحة تاريخية

إن التحليل الدلالي للغات الطبيعية ليس بفكرة مستحدثة من حيث المبدأ فلقد بدأ التفكير به منذ زمن طويل مضى مرّ خلالها باتجاهات ومناهج عديدة من أجل الوصول إلى منهجية مناسبة لطبيعة ومتطلبات اللغات الطبيعية والتي تتيح بدورها إمكانية الوصول إلى تمثيل معرفي لمحتوى تلك اللغات. ولطالما كانت قضية ارتباط التحليل الدلالي بالتحليل النحوي تحتل مرتبة عالية من بين القضايا اللغوية المختلفة. وقد أشار

النحاة العرب إلى هذا الارتباط على سبيل المثال في تعريفهم للفاعل النحوي أنه من قام بالفعل في «كتب محمد» أو من وقع عليه الفعل أحيانا في «سقط محمد».

ويرى كريستوفر بتلر أنه يمكن تمييز ثلاثة مناهج مختلفة للعلاقة بين التحليل الدلالي والتحليل النحوي؛ يقوم أحد هذه المناهج على البدء بالتحليل النحوي، والذي ينتج عنه شجرة نحوية؛ يتبعه تحويل هذه الشجرة النحوية إلى تمثيل دلالي. ولكن لهذا المنهج بعض السلبيات، منها أنه يمثل نموذجا غير معقول للتحليل أو المعالجة التي يقوم بها البشر باعتبار أن الفكرة (المعنى) تنشأ في الذهن أولا ثم يُبنى عليه تركيب الجملة عند إنتاجها، كما أنه لا يقر بإمكانية استخدام المعلومات الدلالية في توجيه التحليل النحوي حيث يمكن للتحليل النحوي أن يكون مسئولا عن إيجاد أكثر من تفسير ممكن؛ وبالطبع يستحيل ذلك إذا كان الانطلاق من النحو. وظهر بوضوح المنهج الثاني الذي أشار إليه كريستوفر باتلر في نهاية الستينيات والسبعينيات؛ ويعتمد على تقليل التحليل النحوي وزيادة التركيز على التحليل الدلالي، وقد تم بناء أنظمة تحليل دلالي تعتمد علي هذا المنهج منها نظام «الإطار النظري للتبعية المفاهيمية» (Conceptual dependency) [٣٤] ونظام آخر يعتمد على ما يسمى بـ «دلالة التفضيل» (Preference seman-tics) [٣٨] وهو نظام لا يتعامل مع القيود الدلالية بين المفاهيم كقيم مطلقة ولكن تبعا لمعايير التفضيل. فعلى سبيل المثال الفعل «يأكل» يتميز فاعله بأنه كائن حي إلا أن الفاعل غير الحي لا يستثنى بشكل مطلق مثل «إن طابعتي تأكل الورق» وفي هذا إشارة للخصائص الدلالية التي لا بد أن تتوفر في المتعلقات الدلالية للأفعال. وقد اقترح كل من بيرتن وودز عام ١٩٧٦ استخدام «شبكات التحول المزيدة» (Augmented Transition Networks) من أجل التحليل الدلالي التي تتميز بالوضوح (Perspicu-ity) والقدرة الإنتاجية العالية (Generative power) والتمثيل الدقيق (Efficiency of representation) والقدرة على معالجة الانتظام في الظواهر اللغوية (Regulari-ties) وأيضا عموميات اللغة (Generalities). أما المنهج الثالث فيعتمد على الدمج بين التحليل النحوي والتحليل الدلالي؛ أي أن يكون هناك تفاعل دائم بينهما وتهدف الأنظمة القائمة عليه إلى منع إقامة التراكيب عديمة النفع أو غير المقبولة دلاليا من خلال السماح لشكل من أشكال التغذية الارتجاعية الدلالية لعملية التحليل النحوي.

وقد درس اللغويون الصوريون إمكانية تمثيل المعنى بعيدًا عن ارتباطه بالتحليل النحوي من خلال منهجية أخرى تعتمد على إمكانية وضع معنى العبارات اللغوية في بنى صورية يطلق عليها تمثيل المعنى ويطلق على بيئة العمل المستخدمة لتوصيف نحو ودلالة هذه التمثيلات لغات تمثيل المعنى حيث تصف عددا من منهجيات التمثيل المعرفي الدلالي منها؛ منهجية «المنطق من الدرجة الأولى» (First–order logic) حيث تحتوى هذه المنهجية على القواعد والأصول اللازمة لصياغة نظريات الذكاء الاصطناعي كما تعتمد على مبادئ المنطق البولياني (Boolean Logic) ومنطق القضايا (Proposi-tional Logic) وتعتبر إحدى منهجيات تمثيل المعرفة التي تمتاز بالمرونة وسهولة الفهم إذ أنها تقدم أساسا حاسوبيا لمتطلبات التحقق والاستنتاج. وكذلك منهجية «التحليل الدلالي القائم على النحو، وتعتمد على مبدأ التركيبية وتكمن فكرتها في أن معنى الجملة يمكن تركيبه من معاني أجزائها بحيث لا يعتمد فقط على معاني الكلمات التي تكونها بل على ترتيب هذه الكلمات في الجملة وطريقة تجميعها والعلاقات فيما بينها وهذا معنى آخر للقول بأن معنى الجملة يعتمد جزئيا على البنية النحوية لها. وبالرغم من الجهود المبذولة منذ القدم للوصول إلى منهج واضح ونظام فعال للتحليل الدلالي إلا أن كل هذه الجهود أسفرت عن مجرد محاولات غير مكتملة وإلى الآن مازال التحليل الدلالي من أصعب مستويات تحليل اللغات الطبيعية.

٢- لغة الشبكات الدلالية الحاسوبية العالمية

إننا كبشر نستخدم اللغة الطبيعية للتعبير عن الحقائق والمعارف. إنّ اللغة الطبيعية مرنة وشاملة بشكل كبير لكنّها تتعدد بتعدد الحضارات وتختلف باختلاف الثقافات (اللغة العربية والإنجليزية والسواحلية ... إلخ) كما أن اللبس جزءً من طبيعتها التي يصعُب عليها التخلص منها لكنها تعالجه بما يسبق العبارات الملتبسة وما يلحقها من نصوص تزيل ذلك اللبس، فيتمكن العقل البشري من فهم الحقائق من خلال سياق الحديث وربطه بما يُحيط بالكلام من ملابسات وظروف وحال المتحدث والمُخاطَب... الخ. ولطالما كان يطمح مجال تمثيل المعرفة إلى إيجاد لغة واضحة وغير مبهمة لتمثيل المعارف ولتكون اللغة المشتركة بين الجنس البشري والآلات، هذه اللغة يجب أن تمكن الحاسوب من التفكير بالمعطيات واستنباط حقائق جديدة من هذه المعطيات ومن

ثم حل المُشكلات المتعلقة بمجال الذكاء الاصطناعيّ. ولعل لغة الشبكات الدلالية الحاسوبية العالمية تمتلك ما يؤهلها لتحقيق هذا الهدف؛ إذ أنها جاءت كمحاولة للتوسط بين الشكل المعرفي المجرد للمحتوى الذي يعبر عنه البشر في حياتهم اليومية وبين الشكل اللغوي الذي يستخدم للتعبير عن هذا المحتوى في شكل جمل وعبارات عن طريق تمثيل المحتوى تمثيلا صحيحا ومتكاملا يختلف عن طريقة تمثيل اللغات الطبيعية له؛ فبينما تقوم اللغات الطبيعية بتمثيل المحتوى في صورة مفردات لغة معينة وتراكيب تتبع قواعد هذه اللغة، فإن لغة الشبكات الدلالية الحاسوبية العالمية لها مفردات وتراكيب تمكنها من تمثيل المحتوى تمثيلا مجردا يحمل كل ما كان يحويه النص الأصلي من معلومات صرفية ونحوية ودلالية وبرجماتية في شكل شبكة دلالية دون انحياز لمفردات أو تراكيب لغة معينة أو حتى مجموعة من اللغات؛ كأن تنحاز لتراكيب اللغة الإنجليزية أو اللغات جرمانية الأصل مثلا. هذا التمثيل الدلالي مكّن لغة الشبكات الدلالية الحاسوبية العالمية من لعب دور اللغة الوسيطة بين اللغات الطبيعية. ويوضح الجدول (٣-١) الفرق بين اللغات الطبيعية ولغة الشبكات الدلالية الحاسوبية العالمية في تمثيل نفس المحتوى حيث يظهر منه أن لغة الشبكات الدلالية الحاسوبية العالمية هي لغة وسيطة بين جميع اللغات الطبيعية، فجميع اللغات الطبيعية في الجدول (٣-١) يمكن الربط بينها باستخدام التمثيل المعرفي للغة الشبكات الدلالية الحاسوبية العالمية.

	اللغات الطبيعية	**لغة الشبكات الدلالية الحاسوبية العالمية**
	العربيـة: أكل الولد التفاحـة	agt(201168468:64.@past.@entry,110285313:59.@def) obj(201168468:64.@past.@entry,107739125:77.@def)
	الإنجليزية: The boy ate the apple	
	الفرنسية: Le garçon a mangé la pomme	

الجدول ٣-١ : تمثيل اللغات الطبيعية ولغة الشبكات الدلالية الحاسوبية العالمية لمحتوى «أكل الولد التفاحة».

وبالرغم من أن لغة الشبكات الدلالية الحاسوبية العالمية هي لغة وسيطة كما ذكرنا إلا أنها تختلف عن اللغات الأخرى التي يطلق عليها لغات وسيطة كالإسبرانتو[1] مثلاً التي هي لغة وسيطة يمكن للبشر استعمالها للتواصل في حياتهم اليومية، لكن لغة الشبكات الدلالية لغة اصطناعية مصممة لمحاكاة التواصل الإنساني وعلى مستوى الآلة.

وقد انطلق برنامج تطوير لغة الشبكات الدلالية الحاسوبية العالمية[2] عام ١٩٩٦ عندما بدأ معهد الدراسات المتقدمة في طوكيو الدعوة لهذا المشروع. وتقوم حاليا «مؤسسة لغة الشبكات الدلالية الحاسوبية العالمية» (UNDL Foundation)[3] بتطوير هذا المشروع وإدارته والإشراف على تنفيذه. وقد اشترك في هذا المشروع حتى الآن سبع عشرة لغة تقوم مؤسساتها على بناء وتطوير الأدوات والموارد اللازمة لتحليل وتوليد هذه اللغات والتي من بينها اللغة الإسبانية والفرنسية واليابانية والبرتغالية والتايلاندية واللغة العربية التي يتم حاليا بناء المكون الخاص بها في مركز اللسانيات الحاسوبية العربية[4] في مكتبة الإسكندرية في مصر.

٣- المكونات اللغوية للغة الشبكات الدلالية الحاسوبية العالمية

لكي تتمكن لغة الشبكات الدلالية الحاسوبية العالمية من محاكاة وظائف اللغات الطبيعية بشكل ناجح كان لا بد أن يكون لها نفس خصائص اللغات الطبيعية ومكوناتها اللغوية من مفردات (UNL Vocabulary). وعلاقات دلالية تربط بين الكلمات (UNL Relations) وتمثل نحو لغة الشبكات الدلالية الحاسوبية العالمية، وهي ما

١- الإسبرانتو لغة مصطنعة سهلة، اخترعها لودفيغ أليعزر زامنهوف كمشروع لغة اتصال دولية عام ١٨٨٧.

٢-جدير بالذكر أن للغة الشبكات الدلالية الحاسوبية إصدارين؛ الأول كان في بداية إطلاقها واستمر لعدة أعوام بعدها خضعت لغة الشبكات الدلالية الحاسوبية لمرحلة تطوير وتحسين نتج عنها إصدار جديد أُطلق عليه UNL+3 هذا الإصدار هو نفسه الذي نتعرض له في هذا الفصل بالشرح والتفصيل ولم نأتي على ذكر الإصدار الأول لكن بالإمكان معرفة المزيد عنه عن طريق هذا الرابط: http://www.undl.org، أما الإصدار الثاني فيمكن متابعته عن طريق هذا الرابط : http://www.unlweb.net/unlweb.

٣- لمعرفة المزيد عن المؤسسة وأنشطتها يُرجى اتباع هذا الرابط:
http://www.undlfoundation.org/undlfoundation/

٤- إسهامات المركز العربي في دعم لغة الشبكات الدلالية الحاسوبية يمكن متابعتها من خلال هذا الرابط الخاص بالمركز العربي في مكتبة الإسكندرية: http://www.bibalex.org/unl/Frontend/home.aspx

يقابل تراكيب الجمل في اللغات الطبيعية من فعل وفاعل ومفعول وتختلف في اللغات الطبيعية من لغة إلى أخرى بحسب نظام اللغة لكنها ثابتة في لغة الشبكات الدلالية الحاسوبية العالمية، وخصائص تحمل المعاني الناتجة عن الظروف غير اللغوية (UNL Attributes). وأنطولوجيا للمفاهيم (UNL Ontology). وهذا ما يتناوله هذا الجزء بالشرح والإيضاح.

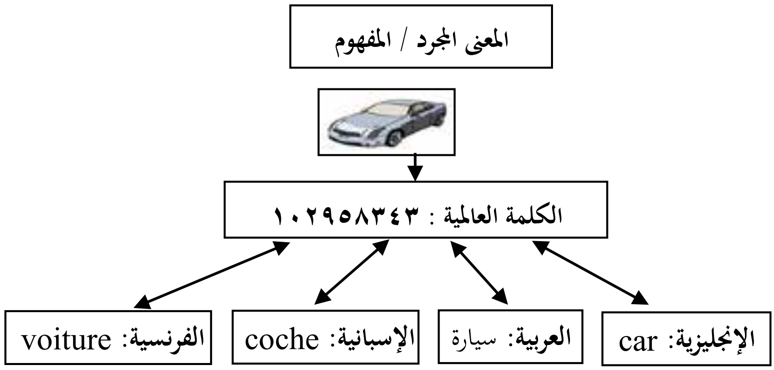
٣,١ – مفردات لغة الشبكات الدلالية الحاسوبية العالمية (UNL Vocabulary)

إنّ المفردة هي وحدة اللغة التي تحمل المعنى، وبالنسبة للغات الطبيعية فإن شكل المفردة يختلف باختلاف اللغة على الرغم من أن المعنى واحد. ولأن لغة الشبكات الدلالية الحاسوبية العالمية لا تنحاز لأي لغة طبيعية فقد قامت بتمثيل المفردة بطريقة مجردة بعيدا عن الشكل المرتبط بلغة بعينها حيث يتيح نظام لغة الشبكات الدلالية الحاسوبية العالمية التعبير عن المعنى دون المبنى في صورة ما يسمى «بالكلمات العالمية» (Universal Words)[١] وهي تمثل المعاني المجردة التي تعبر عن المفاهيم الإنسانية مثلها مثل كلمات اللغة الطبيعية بما فيها من أسماء وأفعال وصفات وأحوال. ويرجع وصف مفردات لغة الشبكات الدلالية بالعالمية إلى أن مدلول هذه المفردات واحد بالنسبة لكل اللغات الطبيعية مثلها مثل إشارة المرور بألوانها الثلاث فكل لون له مدلول ثابت لا يمكن الاختلاف عليه في جميع أنحاء العالم مهما اختلفت اللغات وتعددت الثقافات ومن هنا كان للكلمات العالمية القدرة على التوسط بين جميع اللغات الطبيعية.

ويوضح الشكل (٣-١) مثالاً على ذلك فنفس المفهوم الذي تعبر عنه اللغات المختلفة ببنى مختلفة تقوم لغة الشبكات الدلالية الحاسوبية العالمية بالتعبير عنه بشكل مختلف يتوسط جميع تلك اللغات فيكون بإمكاننا أن نستبدل البنية الإنجليزية أو العربية أو الفرنسية أو الإسبانية للمفهوم بالكلمة العالمية دون انحياز للغة ما أو اختلاف على المعنى الذي تحمله المفردة.

---

١ – لمعرفة المزيد عن الكلمات العالمية برجاء اتباع هذا الرابط:
http://www.unlweb.net/wiki/index.php/Universal_Words



الشَّكل ٣-١: يوضح توسط الكلمة العالمية بين اللغات الطبيعية.

وخلافا لمفردات اللغات الطبيعية فإن مفردات لغة الشبكات الدلالية الحاسوبية العالمية تخلو من أشكال الالتباس، فعلى سبيل المثال كلمة «فصل» في اللغة العربية يمكن استخدامها للتعبير عن أكثر من مفهوم مثل «فصل في كتاب» أو «فصل من فصول السنة» أو «عملية تفريق شيء عن آخر» وغيرها من المعاني المختلفة مما قد يسبب غموض ناتج عن أن نفس المبنى يستخدم للتعبير عن أكثر من معنى وهنا يأتي دور الكلمات العالمية في التعبير عن المعاني المختلفة دون لبس أو غموض حيث يوضح الجدول (٣-٢) طريقة التعبير عن كلمة «فصل» باستخدام مفردات لغة الشبكات الدلالية الحاسوبية العالمية فنفس الكلمة العربية يُقابله أربع كلمات عالمية تعبر عن المعاني المختلفة لهذه الكلمة العربية.

الكلمة العربية	الكلمة العالمية	المعنى
فصل	١٠٦٣٩٦١٤٢	فصل في كتاب (جزء من كتاب).
	١١٥٢٣٦٤٧٥	فصل من فصول السنة الأربعة.
	١٠٣٠٣٨٦٨٥	حجرة دراسية في مدرسة.
	١٠٠٢١٦١٧٤	عملية تفريق شيء عن آخر.

الجدول ٣-٢: التعبير عن المعاني المختلفة لكلمة «فصل» باستخدام لغة الشبكات الدلالية الحاسوبية العالمية.

وقد يتبادر إلى أذهاننا أن الكلمة العالمية هي كلمة مكونة من أحرف كعادة كلمات اللغات الطبيعية، ولكن على خلاف ذلك فإن الكلمة العالمية يتم التعبير عنها برقم كودي. هذا الرقم مأخوذ من شبكة الكلمات الإنجليزية (WordNet) وقد استُخدمت هذه الأرقام للتعبير عن مفردات لغة الشبكات الدلالية الحاسوبية العالمية لعدة أسباب منها أن شبكة الكلمات الإنجليزية عبارة عن شبكة دلالية متكاملة للكلمات الإنجليزية (حوالي ٦٥٩ , ١١٧ مفهوم) وتحتوي على معلومات عن معاني هذه المفاهيم والعلاقات الأنطولوجية بينها وبين المفاهيم الأخرى داخل شبكة المعاني. بالإضافة إلى أن هناك محاولات فعلية من قبل العديد من اللغات منها اللغة الفرنسية والهندية لبناء شبكة كلمات فرنسية (French WordNet) وشبكة كلمات هندية (Hindi WordNet) اعتمادًا على شبكة اللغة الإنجليزية، فإذا تم استخدام نفس الشَّفرات الرقمية الموجودة داخل الشبكة الإنجليزية للتعبير عن الكلمات العالمية في لغة الشبكات الدلالية سيجعل من السهل على كل لغة من اللغات المشاركة في برنامج لغة الشبكات الدلالية بناء شبكة كلمات خاصة بلغتها.

٣ , ٢ - العلاقات الدلالية (UNL Relations)

تهدف لغة الشبكات الدلالية الحاسوبية العالمية إلى بناء شبكة دلالية عالمية لأي جملة طبيعية تعبر عن محتوى تلك الجملة حيث تتشكل تلك الشبكة عن طريق ربط المفردات بعلاقات تعبر عن الدور الدلالي لكل مفردة داخل الجملة مثل علاقات الفاعل الدلالي (Agent) والمفعول الدلالي (Object) والمكان (Place) والزمن (Time) ... وغيرها من العلاقات الدلالية المختلفة التي تربط بين كل كلمتين على حده في الجملة ويتم التعبير عنها برموز مكونة من ثلاثة أحرف مثل: agt (فاعل) وobj (مفعول) وplc (مكان) وtim (زمن) .... الخ. ويبلغ عدد العلاقات الدلالية الموجودة بلغة الشبكات الدلالية الحاسوبية العالمية حوالي ٤٦ علاقة تعبر عن جميع العلاقات الدلالية الممكنة بين كلمات الجمل في أية لغة طبيعية. ومثال على ذلك الجملة العربية (١):

سيلعب الفريق المباراة النهائية في القاهرة يوم الجمعة القادم (١)

فالعلاقات الدلالية التي تربط بين كلمات هذه الجملة هي علاقة agt أو فاعل دلالي بين «سيلعب» و«الفريق» وعلاقة obj أو مفعول دلالي بين «سيلعب» و«المباراة» وعلاقة

plc أو مكان بين «سيلعب» و«في القاهرة» وعلاقة tim أو زمن بين «سيلعب» و«يوم الجمعة». ويختلف التعبير عن أجزاء الجملة بعلاقات دلالية على حسب معنى الجملة فليس كل فاعل نحوي هو فاعل دلالي، فمثلا العلاقة الدلالية بين «لعب» و«محمد» في جملة «لعب محمد» وهي علاقة فاعل دلالي (agt) تختلف عن العلاقة الدلالية بين «انكسر» و«الزجاج» في جملة «انكسر الزجاج» وهي علاقة مفعول دلالي (obj) على الرغم من اتفاق العلاقة النحوية (فاعل نحوي) حيث أن «الزجاج» واقع عليه الفعل وليس قائم به.

٣,٣- السمات (UNL Attributes)[1]

إن الكلام البشري يحمل الكثير من المعاني الضمنية التي لا يمكن التعبير عنها بالكلمات ولكن تفهم من خلال طريقة القول أو تنغيمات الجمل أو نبرة الصوت، فكيف للحاسوب الوصول لهذه المعاني وفهمها والتعامل معها وهي ليست معلومات صرفية ولا نحوية ولكنها تتضح من سياق الكلام وطريقة التعبير كما أنها مهمة لنقل المعنى السليم وقد تؤدي لاختلاف معاني الجمل مثل جملة «أأتممت عملك اليوم» يمكن أن تعبر عن استفهام أو عن استنكار وفقا لما يقصده المتكلم. وقد وضعت لغة الشبكات الدلالية الحاسوبية العالمية طريقة للتعبير عن هذه المعلومات عن طريق مجموعة من الرموز الإضافية (السمات) التي تستخدم في التحليل الدلالي لإضافة المعلومات التي لم يتم التعبير عنها بمفردات لغة الشبكات الدلالية أو بالعلاقات الدلالية إذ إنَّها تُستَخدَم للتعبير عن ثلاثة أنواع من المعلومات؛ أولاً: معلومات عن دور المفهوم داخل الشبكة الدلالية مثل السمة «@entry» ومعناها «المدخل للشبكة الدلالية» وتمثل المفهوم الأساسي (Main Predicate) وتوضع لتوضح الكلمة التي تمثل «مدخل» الشبكة الدلالية، هذا المدخل ترتبط به المفردات الرئيسية داخل الشبكة الدلالية بشكل مباشر والمفردات الأخرى بشكل غير مباشر، ويُعد بمثابة مفتاح الشبكة الدلالية. ثانيًا: معلومات مورفولوجية مثل التي تحملها السوابق واللواحق كالزمن والتذكير والتأنيث والعدد وغيرها من المعلومات فعلى سبيل المثال لتمثيل الفعل

١-لمعرفة المزيد عن السمات في لغة الشبكات الدلالية الحاسوبية يُرجى اتباع هذا الرابط:
http://www.unlweb.net/wiki/index.php/Attributes

«يكتب» تستخدم السمة «present@» للتعبير عن الزمن المضارع وتوضع على الرقم الكودي للمفهوم «كتب» مثل «200993014 .present@»، والسمتين «def@» و «indef@» لتحديد التعريف والتنكير للكلمات والسمتين «female@» و«male@» للتعبير عن التأنيث والتذكير للكلمة كما في «male@.110020890» و« female@.110020890». ثالثًا: معلومات خاصة بالسياق مثل السمة polite@ والتي تصف التهذيب في عبارة «وتفضلوا سيادتكم» و exclamation@ والتي تعبر عن التعجب كما في جملة «يا له من منظر رائع» وغيرها من السمات الأخرى.

٣,٤ – الأنطولوجيا اللغوية (UNL Ontology)[1]

لكي تكتمل المكونات اللغوية للغة الشبكات الدلالية كان لا بد لها أن يكون لديها مكون آخر يشبه مكونات اللغات الطبيعية التي تمثل القدرة اللغوية للإنسان ألا وهو الأنطولوجيا وقاعدة المعرفة. وأنطولوجيا لغة الشبكات الدلالية هي منظومة الكلمات العالمية ولكن في بناء شجري تُرتب فيه الكلمات العالمية بشكل هرمي طبقا للعلاقات الأنطولوجية بينها. هذه العلاقات تعبر عن ارتباط الكلمات العالمية مع بعضها البعض بعلاقات هرمية مثل علاقة «نوع من» (icl) وعلاقة «مثال ل» (iof). والمثال (٢) يوضح شكلَ أحد مداخل الأنطولوجيا. وهو يعني أن التفاحة وهي ما يعبر عنها بالكلمة العالمية الأولى من جهة اليسار (١٠٧٧٣٩١٢٥) هي نوع من الفاكهة وهي ما يعبر عنها بالكلمة العالمية الثانية من جهة اليسار (١١٣١٣٤٩٤٧)، ويعبر عن هذا باستخدام علاقة (icl) أما الرقم ١ في النهاية يعبر عن صحة هذه العلاقة الأنطولوجية بين الكلمتين العالميتين.

(٢)	icl(<[[107739125]];[[113134947]])=1;

ولعل ذلك يوضح أهمية الأنطولوجيا في لغة الشبكات الدلالية الحاسوبية العالمية حيث أن بهذه المنهجية يمكن للغة الهدف التعرف على معنى مفهوم ما مرتبط بثقافة اللغة المصدر (في حالة الترجمة). فمثلا يمكن للغات الأخرى التعرف على مفهوم كلمة

١ – يمكن معرفة المزيد عن الأنطولوجيا اللغوية عن طريق هذا الرابط:
http://www.unlweb.net/wiki/index.php/Ontology

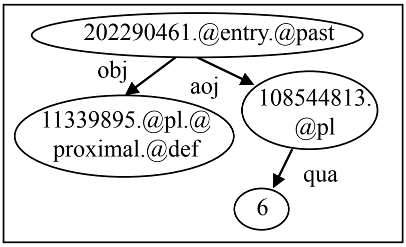
«الإحرام» في العربية بالرغم من عدم وجود البعد الثقافي لهذا المفهوم في تلك اللغات.

ويرتبط بالأنطولوجيا ما يسمى بقاعدة معرفة لغة الشبكات الدلالية الحاسوبية العالمية (UNL Knowledge Base) وهي تختلف عن الأنطولوجيا من حيث نوع وطبيعة العلاقات فتضم قاعدة المعرفة شبكة من الكلمات العالمية تربط بينها علاقات لغة الشبكات الدلالية بينما تحتوي الأنطولوجيا على علاقات هرمية فقط كما هو موضح في المثال (٢). وعلى هذا فإن قاعدة المعرفة تشتمل على الأنطولوجيا ولكنها أعم وأشمل كما يتضح ذلك من المثال (٣). ففي المثال (٣) (أ) نجد نفس العلاقة الهرمية بين الكلمتين العالميتين (١٠٠٠٠١٩٣٠) و(١٠٠٠٠١٧٤٠) كما في المثال (٢)، وبالإضافة إلى ذلك نجد نوعا آخرا من العلاقات كما في ٣ (ب) حيث لا يمكن تحقق العطف بين علاقتين إحداهما (agt) والأخرى ليست (agt)، فإذا كان هناك x و y بينهما علاقة (agt) و x و y ليس بينهم علاقة (agt) لا يمكن العطف بينهم. ويدل الرقم ٠ في نهاية المدخل على عدم تحقق العلاقة.

أ- icl(<[[100001930]];[[100001740]])=1; ب- and(agt(x;y);^agt(x;y))=0;	(٣)

وفي ختام هذا العرض للمكونات اللغوية للغة الشبكات الدلالية الحاسوبية العالمية، يوضح الشكل (٣-٢) الشبكة الدلالية الحاسوبية للجملة العربية في المثال رقم (٤) التي تم تمثيلها باستخدام المفردات العالمية والعلاقات الدلالية والسمات لتشترك جميعا للتعبير عن معنى الجملة. وقد أمكن التعبير عن المعنى المجرد لهذه الجملة دون الشكل المقيد بلغة ويمكن لأي لغة طبيعية فهم هذا التمثيل وذلك يجعلنا نشعر بدقة معنى العالمية في لغة الشبكات الدلالية الحاسوبية العالمية. ويمكن قراءة الشبكة الدلالية بدءً من المفهوم الذي يمثل المدخل والذي يحمل السمة «entry@» فهو المفهوم الأساسي للشبكة الدلالية. أما الشكل (٣-٣) فيعبر عن الجملة العربية مكتوبة بلغة الشبكات الدلالية الحاسوبية العالمية. والجدول رقم (٣-٣) يوضح المقابل العربي لكل مفهوم من مفاهيم تلك الشبكة الدلالية.

استفادت من هذه القروض ست دول (٤)



الشَّكل ٣-٢: الشبكة الدلالية الحاسوبية لجملة العربية.

المفهوم بلغة الشبكات الدلالية	المقابل باللغة العربية
٢٠٢٢٩٠٤٦١	استفاد
١٠٨٥٤٤٨١٣	دول
١١٣٣٩٨٩٥٣	قرض

الجدول ٣-٣: المقابل باللغة العربية لمفاهيم الشبكة الدلالية.

```
{unl}
aoj(202290461:00.@past.@entry,108544813:1Z.@pl)
obj(202290461:00.@past.@entry,113398953:0Y.@pl.@proximal.@def)
qua(108544813:1Z.@pl,6:1N)
{/unl}
```

الشَّكل ٣-٣: الجملة العربية مكتوبة بلغة الشبكات الدلالية الحاسوبية العالمية.

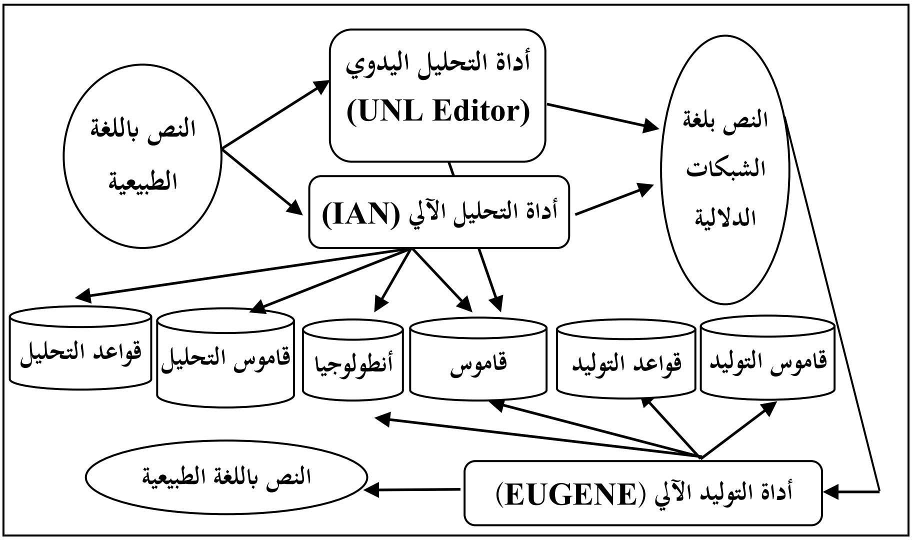
٤- موارد وأدوات لغة الشبكات الدلالية الحاسوبية العالمية

بالإضافة إلى ما تم ذكره عن المكونات اللغوية للغة الشبكات الدلالية الحاسوبية العالمية نتعرض في هذا الجزء لموارد وأدوات لغة الشبكات الدلالية الحاسوبية العالمية والمتمثلة في قواميس (Dictionaries)، وقواعد صورية (Formal Rules)، ومحركات لغوية (Engines) والتي تتلخص مهمتها في تحليل النصوص الواردة من اللغات الطبيعية وتمثيل معناها في شكل شبكة دلالية والعكس، أي فك الشبكة الدلالية إلى أي لغة طبيعية في شكل جملة صحيحة تتماشى مع مفردات وتراكيب اللغة الهدف سواء كانت إنجليزية أو عربية أو فرنسية ... الخ.

والشكل (٣-٤) يعبر عن رسم توضيحي لآلية عمل موارد وأدوات لغة الشبكات الدلالية الحاسوبية العالمية. فتقوم أدوات التحليل بتحليل النص المكتوب باللغة الطبيعية باستخدام موارد اللغة المصدر (قواعد التحليل وقاموس التحليل) بالإضافة إلى موارد لغة الشبكات الدلالية (قاموس لغة الشبكات الدلالية UNL) والذي يحتوي

على الكلمات العالمية وما تحتاجه من خصائص ومعلومات خاصة بكل كلمة وكذلك الأنطولوجيا الخاصة بلغة الشبكات الدلالية ليخرج النص المكتوب باللغة الطبيعية في صورة شبكة دلالية.

وتقوم أدوات التوليد بإعادة فك التمثيل الدلالي المتمثل في النصوص المكتوبة بلغة الشبكات الدلالية إلى أي لغة طبيعية مطلوبة باستخدام موارد اللغة الهدف (قاموس التوليد وقواعد التوليد) وكذلك موارد لغة التواصل العالمية أيضا (قاموس UNL والأنطولوجيا الخاصة بلغة الشبكات الدلالية لتخرج في النهاية الجملة باللغة الطبيعية.



الشَّكل ٣-٤: آلية عمل موارد وأدوات لغة الشبكات الدلالية الحاسوبية العالمية.

٤ ,١- قاموس لغة الشبكات الدلالية الحاسوبية العالمية[١]

يُعد القاموس بمثابة القلب الذي يضخ لأي نظام لغوي ما يلزمه من معلومات لكي يعمل بشكل جيد وفعال لذلك يجب أن تتوافر فيه المواصفات اللغوية التي تمكنه من أداء هذا الدور. وقاموس لغة الشبكات الدلالية الحاسوبية العالمية هو حجر الزاوية في عمليتي التحليل والتوليد والذي تتوافر فيه كل المعلومات اللازمة للوصول إلى الشبكة

١- لمعرفة المزيد عن قاموس لغة الشبكات الدلالية الحاسوبية يرجى اتباع هذا الرابط:
http://www.unlweb.net/wiki/index.php/Dictionary_Specs

الدلالية الناتجة عن التحليل الدلالي للجملة الطبيعية وكذلك الوصول للجملة الطبيعية الناتجة عن التوليد الآلي لتلك الشبكة الدلالية حيث أنه يحتوي على نوعين رئيسيين من المعلومات. الأول: الكلمة في اللغة الطبيعية وما يقابلها في لغة الشبكات الدلالية الحاسوبية العالمية إذ أنها وسيلة الربط بين كلمات اللغات الطبيعية والكلمات العالمية. والثاني: معلومات لغوية تصف السلوك اللغوي لكلمات اللغة الطبيعية.

ويوضح الشكل (٣-٥) صُورةَ المدخل القاموسي الخاص بلغة الشبكات الدلالية الحاسوبية العالمية ومكوناته. فتحتوي خانة الـ [NLW] على الكلمة باللغة الطبيعية وخانة الـ «UW» على مقابلها في لغة الشبكات الدلالية الحاسوبية العالمية وخانة الـ (ATTR) على مجموعة الخصائص اللغوية التي توضع مع كل كلمة لوصف سلوكها اللغوي. أما الخانة الأخيرة من القاموس فتحتوي على ثلاث أنواع من المعلومات: أولا (FLG) وتعبر عن لغة القاموس سواء كانت عربية أو إنجليزية أو غيرها من اللغات الطبيعية. ثانيا (FRE) وتعبر عن تكرار ظهور الكلمة داخل اللغة الطبيعية وتفيد في عملية التحليل. ثالثا (PRI) وتعبر عن أولوية استخدام الكلمة في اللغة الطبيعية وتفيد في عملية التوليد.

[NLW] {ID} "UW" (ATTR, ...) <FLG, FRE, PRI>;

الشَّكل ٣-٥: الشكل العام لمدخل القاموس الخاص بلغة الشبكات الدلالية الحاسوبية العالمية.

ويتيح القاموس إمكانية تخزين كل أنواع الكلمات سواء كانت بسيطة أو مركبة أو كلمة متعددة المفاهيم، فعلى سبيل المثال يتيح قاموس لغة الشبكات الدلالية الحاسوبية العالمية إمكانية تخزين الكلمات البسيطة في اللغة العربية مثل كلمة «كتاب» أو الكلمات المركبة مثل «أخذ في الاعتبار» وكذلك الكلمات متعددة المفاهيم مثل «جمهورية مصر العربية». فيوضع بجوار كل كلمة عربية ما يعبر عنها من كلمة عالمية في خانة الكلمة العالمية.

- **قاموس لغة الشبكات الدلالية الحاسوبية العالمية الخاص باللغة العربية**

في إطار لغة الشبكات الدلالية الحاسوبية العالمية يتم تصميم القاموس العربي بحيث يكون المصدر الرئيسي للمعلومات التي تتطلبها عمليتي تحليل وتوليد النصوص من وإلى العربية إذ أنه وسيلة الربط بين مفردات اللغة العربية ومفردات لغة الشبكات الدلالية

الحاسوبية العالمية والتي تمنع حدوث أي لبس أو غموض في معنى الكلمة العربية فكل مدخل في القاموس يعبر عن مفهوم واحد لكلمة عربية محددة، وقد تتكرر الكلمة العربية (من حيث المبنى) في أكثر من مدخل لكن معناها يكون مختلف فيتم تمثيل كل معنى من تلك المعاني بكود رقمي مختلف. ويحتوي القاموس العربي على كل المعلومات اللغوية الخاصة بالكلمة العربية والتي تصف السلوك اللغوي للكلمة صرفيا ونحويا ودلاليا في السياقات المختلفة الأمر الذي يساعد على إتمام عمليتي التوليد والتحليل بنجاح. هذه الخصائص اللغوية نوعان؛ النوع الأول يصف قسم الكلمة إذا كانت اسم، فعل، صفة، ظرف، أداة، سابقة أو لاحقة وغيرها من أقسام الكلام، وتركيب الكلمة إذا كانت كلمة بسيطة أو مركبة أو مفهوم متعدد الكلمات. كما يحتوي على بعض من المعلومات التي تنقسم إلى: معلومات صرفية (مثل الأبواب التصريفية، التجرد، الزيادة، الصحة، الاعتلال، التذكير، التأنيث، الإفراد، التثنية، الجمع...إلخ). ومعلومات نحوية (مثل الصيغة، الزمان، اللزوم، التعدي، التمام، النقصان، البناء للمعلوم، البناء للمجهول،... إلخ). ومعلومات دلالية (مثل الإدراك، الامتلاك، الحركة، الشك، التواصل، التنافس، المشاركة، العاقل، غير العاقل، الوقت، الحالة، العلاقة ...إلخ). على سبيل المثال عند إدراج الفعل «أعطى» داخل القاموس العربي تتم إضافة المعلومات اللغوية التالية له: فعل مزيد - متصرف - معتل الآخر ناقص- يتبع الباب التصريفي «أَفْعَلَ- يُفْعِلُ» وهي معلومات صرفية، فعل ماضي -مبني للمعلوم - متعدي لمفعولين وهي معلومات نحوية، وأنه فعل حركي وهي معلومة دلالية، وأخيرًا أن احتمال ظهوره أعلى من أفعال أخرى مثل «منح» وهي معلومة إحصائية.

أما النوع الثاني من المعلومات اللغوية فيصف سلوك الكلمة في السياقات والتراكيب المختلفة وينقسم إلى قسمين، القسم الأول مسئول عن اشتقاق الأشكال التصريفية المختلفة للكلمات العربية. فاللغة العربية كما نعرف غنية بالاشتقاقات والكلمة الواحدة ينتج عنها عدد كبير من التصريفات المختلفة وفقًا للسياق الواردة فيه لذلك كان لا بد من وضع معلومة تختص باشتقاق الكلمات تسمى القوالب الصرفية حيث يمكن لهذه القوالب أن تتعامل مع كافة الكلمات العربية أيًا كانت فئتها المعجمية (الأسماء، الأفعال، الصفات، الظروف) وذلك بمراعاة العوامل والمعايير المختلفة التي تؤثر على كل فئة منها. فمثلا الفعل «أعطى» فعل مزيد ومتعدي وناقص يتبع الميزان الصرفي

«أَفْعَلَ - يُفْعِلُ» وبتطبيق القالب الصرفي الخاص بهذه النوعية من الأفعال يتم توليد الأشكال الصرفية المختلفة للفعل «أعطى» وهي: (يعطي - يعطى - أعطيا - يعطيان - يعطيا - يعطون - يعطوا - تعطين - أعطي - يعطين - أعطينا - نعطي). ونفس الحال بالنسبة لأسماء اللغة العربية حيث تتمكن القوالب الصرفية الخاصة بها من اشتقاق الأشكال المختلفة للجموع المنتظمة وغير المنتظمة وتلك التي تعبر عن المثنى فعلى سبيل المثال الاسم «بريء» تتمكن القوالب الصرفية من اشتقاق الأشكال الصرفية المختلفة له وهي: «بريئة - أبرياء - بريئان - بريئين - بريئات - بريئتان - بريئتين». وغيرها من الكلمات والأشكال المختلفة. أما القسم الثاني لهذا النوع من المعلومات فهو مسئول عن وصف السلوك النحوي للكلمة وتحديد عدد ونوع المتعلقات النحوية اللازمة لتلك الكلمة (Subcategorization Frame). ويُطلق عليه القالب النحوي الذي يحدد مواصفات السياق الذي يستخدم فيه الفعل. فعلى سبيل المثال الفعل «أعطى» توضع بجواره معلومات تدل على عدد ونوع المتعلقات النحوية الخاصة به وهي عبارة عن ثلاث متعلقات نحوية: (مخصص الفعل (Verb Specifier (VS) و (المتمم الأول للفعل (Verb Complement (VC) و (المتمم الثاني للفعل (Verb Complement (VC) والتي تظهر واضحة من خلال جملة «الإسلام أعطى للمرأة حقوقها كاملة» فمخصص الفعل (VS) وهو عبارة عن المركب الاسمي (الإسلام) والمتمم الأول للفعل (VC) وهو عبارة عن المركب الاسمي (حقوقها) والمتمم الثاني للفعل (VC) وهو عبارة عن شبه جملة تبدأ بحرف الجر «ل» متبوع بالمركب الاسمي «المرأة». ويوضح الشكل (٣-٦) مثالًا لشكل مداخل القاموس العربي للفعل «أعطى».

[أعطى] {} «200878876« (V,CMV,VER,WRD,TST2,Y18, M222)<ar,0,2>;

الشَّكل ٣-٦: أحد مداخل القاموس العربي للغة الشبكات الدلالية الحاسوبية العالمية.

ويوضح الشكل (٣-٦) صُورةَ الفعل «أعطي» داخل القاموس العربي والمعلومات اللغوية المخزنة بجانبه مثل CMV وهي تعبر عن التصنيف الدلالي للفعل وهو فعل تواصل (communication verb) وتركيب الفعل WRD وهو فعل بسيط، و TST2 وتعني أن الفعل «أعطي» له فاعل وكذلك متعدٍ لمفعولين، و Y18 وهي المعلومة الخاصة بالسلوك النحوي للفعل داخل اللغة العربية والتي تحدد أن الفعل «أعطى» له

ثلاثة متعلقات دلالية، M٢٢٢ وتعبر عن السلوك الاشتقاقي للفعل «أعطى» وتسمح باشتقاق جميع الأشكال المطلوبة لهذا الفعل.

٤ , ٢- التحليل الآلي باستخدام لغة الشبكات الدلالية الحاسوبية العالمية

إن عملية التحليل الآلي باستخدام لغة الشبكات الدلالية الحاسوبية العالمية جعلت التعامل مع الجملة الطبيعية أكثر سهولة ويسرًا إذ أنها تقوم بتمثيل كل ما يمكن أن تحتويه الجملة الطبيعية من معلومات صرفية ونحوية ودلالية وبرجماتية في شكل شبكة دلالية توضح المعنى الدقيق لكل كلمة في الجملة الواردة وماهية العلاقات الدلالية التي تربط كلمات الجملة بعضها بعضا عن طريق استخدام مفردات لغة الشبكات الدلالية (Universal Words) وربطها بعلاقات دلالية (Semantic Relations)، ثم تستعين بالسمات (Attributes) لإضافة المعلومات التي لم يتم التعبير عنها سواء بالمفردات أو العلاقات الدلالية. وتتم عملية التحليل الآلي عن طريق المحلل التفاعلي (Interactive Analyzer- IAN)[1] والتي تستخدم موارد لغة الشبكات الدلالية الحاسوبية العالمية متمثلةً في القاموس الذي أشرنا إليه في الجزء الخاص بقاموس لغة الشبكات الدلالية الحاسوبية العالمية وقواعد التحليل الخاصة بلغة الشبكات الدلالية الحاسوبية العالمية والتي تشتمل على ستة مراحل تبدأ بالمرحلة الأولى وهي تحليل الجملة الطبيعية للتعرف على معاني المفردات من خلال القاموس تليها المرحلة الثانية وهي التحليل الصرفي لكلمات الجملة الطبيعية وتحديد السمات الخاصة بكل كلمة. ثم المرحلة الثالثة وهي بناء العلاقات النحوية بين كلمات الجملة الطبيعية في شكل شجرة نحوية في إطار علم اللغة الحديث. ثم المرحلة الرابعة وهي الانتقال من البنية السطحية للشجرة النحوية إلى البنية العميقة. ثم المرحلة الخامسة وهي تحويل الشجرة النحوية إلى شبكة دلالية. وأخيرًا المرحلة السادسة وهي تنقيح الشبكة الدلالية بعد معالجتها آليا.

١- يُمكن استخدام أداة التحليل الآلي للغة الشبكات الدلالية الحاسوبية عن طريق هذا الرابط:
http://dev.undlfoundation.org/analysis/index.jsp

▪ التحليل الآلي للجملة العربية باستخدام لغة الشبكات الدلالية الحاسوبية العالمية

نتعرض في هذا الجزء بالتوضيح لمراحل التحليل الآلي الست للغة الشبكات الدلالية الحاسوبية العالمية من خلال تطبيقها على الجملة العربية رقم (٥) وحتى نصل إلى الشبكة الدلالية وهي الهدف من هذه العملية.

بدأت جميع الدول تعتمد على الإداريين المتدربين	(٥)

• التعرف على معاني المفردات من خلال القاموس

يبدأ التحليل الآلي بمرحلة التعرف على المفردات العربية واستبدالها بالمفاهيم العالمية حيث تمر الجملة على قاموس لغة الشبكات الدلالية لإيجاد المعنى المقابل لكل كلمة في الجملة كما يظهر في الشكل (٣-٧).

```
[108168978 "{}[دول" (N, PLR) <ara,125,1>;
 [202379528 "{}[بدأت" (V, ICP,FEM,SNG,3PS,PAS) <ara,46,2>;
[202664017 "{}[تعتمد على" (V, MCL,SNG,NOM,2PS,PRS) <ara,46,2>;
[110069645 "{}[إداريين" (N, PLR) <ara,3,1>;
[301911683 "{}[متدرب" (J, ADJ) <ara,0,0>;
```

الشَّكل ٣-٧: ناتج مرحلة التعرف على معاني المفردات.

وبالتالي تتحول المفردة العربية إلى مفهوم من مفاهيم لغة الشبكات الدلالية الحاسوبية العالمية كما في الشكل (٣-٨) والذي نلاحظ منه أن بعض مفردات الجملة لم تُستبدل مثل كلمة «جميع» وأدوات التعريف «ال» وذلك لأنها ليست مُدرجة بالقاموس الخاص باللغة العربية والتعامل معها يتم بمراحل لاحقة وليس بتلك المرحلة.

جميع ال١٠٨١٦٨٩٧٨ ٢٠٢٣٧٩٥٢٨ ٢٠٢٦٦٤٠١٧ ال١١٠٠٦٩٦٤٥ ال٣٠١٩١١٦٨٣

الشَّكل ٣-٨: شكل الجملة العربية بعد مرورها بمرحلة التعرف على المفردات العربية.

• التحليل الصرفي ووسم الكلمات

تبدأ القواعد اللغوية في هذه المرحلة بالتحليل الصرفي للسوابق واللواحق المتصلة ببعض الكلمات والتي قد يتم حذف بعضها واستبدالها بالسمات التي تعبر عن معناها. وبالنسبة للجملة (٥) يتم التحليل الصرفي كما يلي: جميع الكلمات التي تحمل الصفة

«PLR» الدالة على الجمع والمستمدة من القاموس يتم وسمها بالخاصية (@pl) المعبرة عن الجمع. والكلمات التي يتصل بها السابق «ال» يتم وسمها بالخاصية (@def) الدالة على التعريف. أما الأفعال التي تحمل صفة المضارعة «PRS» يتم وسمها بالخاصية (@present). ويوجد في تلك المرحلة نوع آخر من الكلمات يتم استبدالها بسمات لغوية مثل الفعل «بدأت» والذي لديه في القاموس الصفة الدلالية «ICP» الدالة على بدء حدث آخر في الجملة، فيُحذف هذا الفعل وتحل محله الخاصية (@inceptive) والتي توضع على الفعل الذي يليه وهو (تعتمد). وكذلك كلمة «جميع» الدالة على الكلية تُستبدل بالسمة (@all) وتوضع على الكلمة التي تليها وهي «دول» فيكون الشكل الناتج عن تلك المرحلة كما في الشكل (٣-٩):

**108168978.@pl.@all.@def 202664017 @inceptiv. @present.
110069645@pl.@def.301911683**

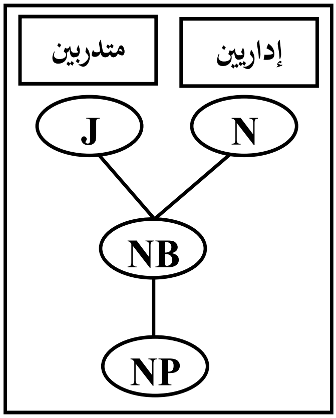
الشَّكل ٣-٩: ناتج مرحلة التحليل الصرفي ووسم الكلمات.

- **بناء العلاقات النحوية بين الكلمات (البنية السطحية للجملة)**

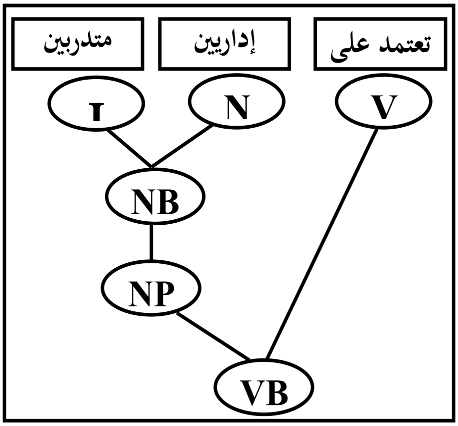
في هذه المرحلة يتم تحويل الجملة العربية المحللة صرفيا إلى شجرة نحوية تُعبر عن البنية السطحية للجملة طبقًا لنظرية (X-Bar) والتي تعرض لها الباب الخامس من هذا الكتاب بمزيد من الإيضاح. فبالنسبة للجملة العربية رقم (٥) نقوم بتطبيق العمليات التالية عليها:

- الاسم «إداريين» ذو الخاصية (N) يتم ربطه بالصفة «متدربين» ذات الخاصية (J) ومن ثم يتم بناء المركب الاسمي الوسيط ((N-Bar (NB). ولأن «إداريين» موسومة بالسمة (@def) - أي أن أداة التعريف التي تعوض عنها تلك السمة تُعد بمثابة مخصص اسمي ((Noun Specifier (NS) للاسم «إداريين» - يتم تحويل المركب الاسمي الوسيط (NB) إلى المركب الاسمي النهائي ((Noun Phrase (NP) كما هو واضح في الشكل (٣-١٠).

- يتم ربط المركب الاسمي النهائي «إداريين متدربيين» (NP) الذي تم بناؤه مع الفعل «تعتمد على» -حيث أن «إداريين متدربين» بمثابة المفعول به بالنسبة للفعل «تعتمد على» والمتممة له ((Verb Complement (VC) - لينشأ المركب الفعلي الوسيط ((V-Bar (VB) كما هو واضح في الشكل (٣-١١).



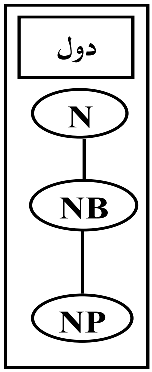
الشكل ٣-١٠: مركب اسمي نهائي.



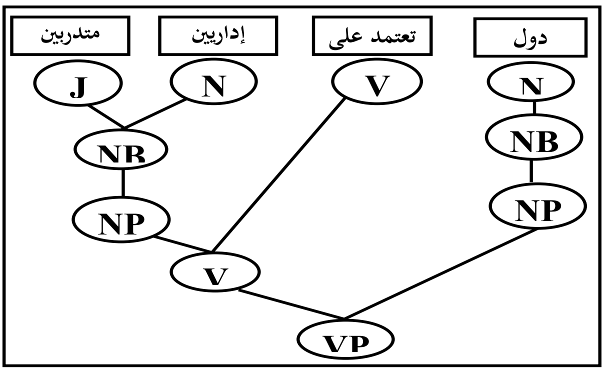
الشكل ٣-١١: مركب فعلي وسيط.

- بعد ذلك يتم تحويل الاسم «دول» (N) والذي هو مركب اسمي وسيط (NB) موسوم بالسمة (@def) – أي أن أداة التعريف التي تعوض عنها تلك السمة تُعد بمثابة مخصص اسمي (NS) للاسم «دول» – إلى مركب اسمي نهائي (NP) كما هو واضح في الشكل (٣-١٢).

- وأخيرًا، يتم ربط المركب الاسمي النهائي «دول @def» (NP) الَّذي تم بناؤه من قبل – والذي يمثل الفاعل بالنسبة للفعل (تعتمد على) والمخصص الفعلي له (verb Specifier– VS) – مع المركب الفعلي الوسيط «تعتمد على إداريين متدربين» (VB) وذلك لبناء التركيب الفعلي النهائي (Verb-Phrase – VP) كما هو واضح في الشكل (٣-١٣).



الشَّكل ٣-١٢: مركب اسمي نهائي.

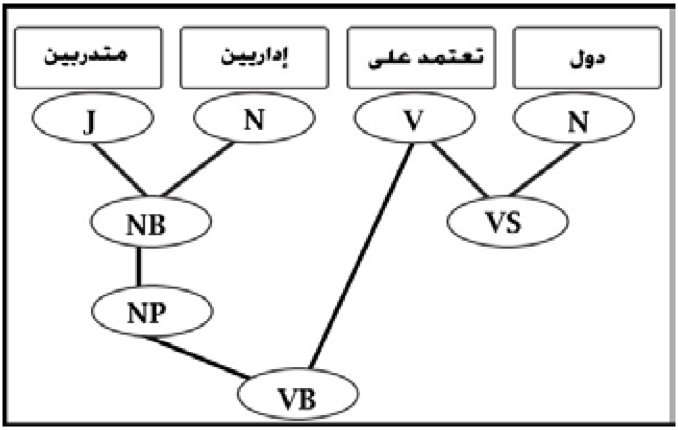


الشَّكل ٣-١٣: مركب فعلي نهائي.

▪ **الانتقال من البنية السطحية للشجرة النحوية إلى البنية العميقة**

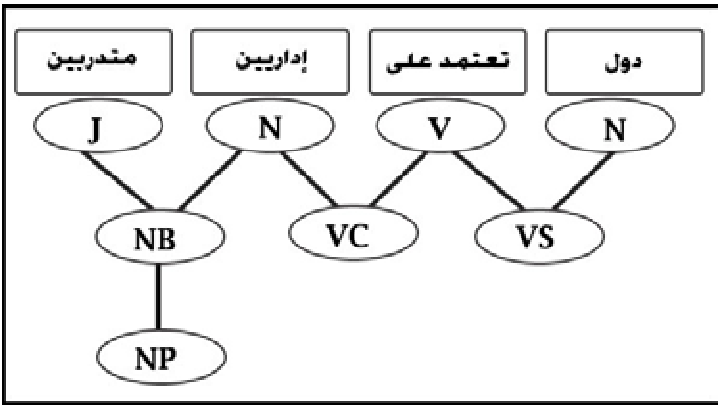
في هذه المرحلة يتم فك الشجرة النحوية التي تم تكوينها في المرحلة السابقة لفروع ثنائية نحوية أكثر تعقيدا على النحو التالي:

- فك التركيب الفعلي النهائي (VP) إلى مركب فعلي وسيط (VB) وبناء علاقة مخصص فعلي (VS) بين الفعل «يعتمد على» والاسم «دول» كما هو واضح في الشكل (٣-١٤).



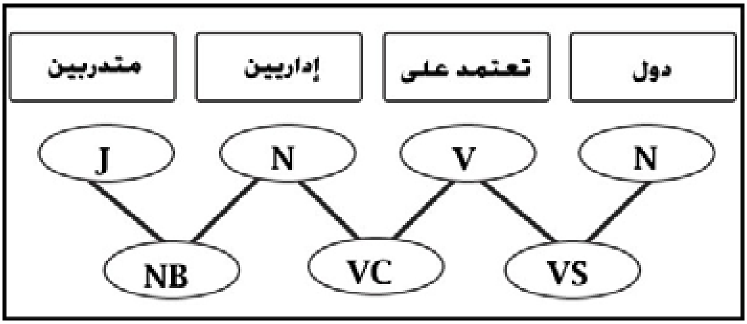
الشَّكل ٣-١٤: بناء المخصص الفعلي (VS).

- يتم تحويل المركب الفعلي الوسيط (VB) إلى العلاقة النحوية متمم فعلي (VC) بين الفعل «يعتمد على» والاسم «إداريين» كما هو واضح في الشكل (٣-١٥).



الشَّكل ٣-١٥: بناء المتمم الفعلي (VC).

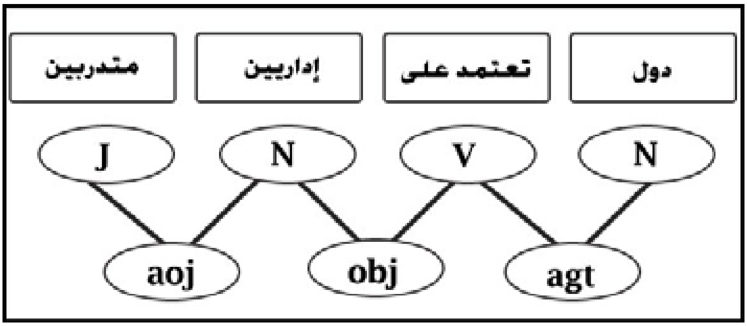
- وأخيرًا يتم تحويل المركب الاسمي النهائي (NP) بين الاسم «إداريين» والصفة «متدربين» إلى العلاقة النحوية ملحق اسمي (Noun Adjunct (NA)) كما هو واضح في الشكل (٣-١٦).



الشَّكل ٣-١٦: بناء الملحق الاسمي (NA).

- **المرحلة الخامسة: تحويل الشجرة النحوية إلى شبكة دلالية للجملة العربية**

في هذه المرحلة يتم تحويل الشجرة النحوية إلى شبكة دلالية معبرة عن محتوى الجملة العربية، حيث يتم تحويل (الملحق الاسمي NA) إلى علاقة (الوصفية الدلالية aoj). وتحويل (المتمم الفعلي VC) إلى علاقة (المفعولية الدلالية obj). وأخيرًا تحويل (المخصص الفعلي VS) إلى علاقة (الفاعلية الدلالية agt) كما موضح في الشكل (٣-١٧).



الشَّكل ٣-١٧: تحويل العلاقات النحوية إلى علاقات دلالية.

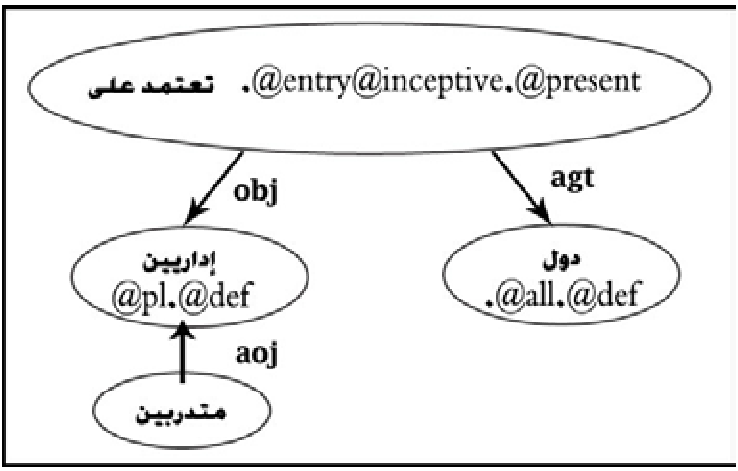
- **تعديل الشبكة الدلالية**

هي مرحلة يتم فيها تعديل الشبكة الدلالية الناتجة عن المراحل السابقة من حيث دمج علاقة أو تقسيمها أو إضافة أخرى... إلخ لكنها مرحلة اختيارية لسنا بحاجة إليها في هذه الجملة. لكن ربما تكون ذات أهمية في حالات أخرى. بالتالي وبعد المرور بالمراحل السابقة جميعها تصبح الجملة العربية ممثلة دلاليا كما في الشكل (٣-١٨).

```
{org}
بدأت جميع الدول تعتمد على الإداريين المتدربين
{/org}
{unl}
agt(202664017:14.@entry.@inceptive.@present,108168978:45.@all.@def)
obj(202664017:14.@entry.@inceptive.@present,110069645:93.@pl.@def)
aoj(301911683:99,110069645:93.@pl.@def)
{/unl}
[/S]
```

الشَّكل ٣-١٨: ناتج أداة التحليل الآلي لجملة عربية.

وهذا التمثيل الدلالي يمكن التعبير عنه في شكل شبكة دلالية سهلة القراءة كما في الشكل (٣-١٩):



الشَّكل ٣-١٩: شبكة دلالية لجملة عربية.

٤ , ٣- التوليد الآلي باستخدام لغة الشبكات الدلالية الحاسوبية العالمية

تحدثنا في الجزء السابق عن التحليل الآلي للجملة الطبيعية وانتهينا بتمثيل جملة عربية في شكل شبكة دلالية ترتبط مفرداتها بعلاقات دلالية. وسنقوم في هذا الجزء بتوضيح كيفية توليد الشبكة الدلالية في شكل جملة طبيعية واضحة المعنى ومتكاملة الأركان وفقًا لقواعد كل لغة. ولكن لا بد في البداية من توضيح مفهوم التوليد الآلي للجملة الطبيعية.

- **التوليد الآلي للجملة الطبيعية**

هو القدرة على بناء مجموعة غير متناهية من الجمل الصحيحة بلغة طبيعية من تمثيل آلي ويهتم التوليد الآلي بالجملة من ثلاث جوانب:

- جانب المعنى: حيث يجب أن تتسم الجملة المولّدة آليًّا بوضوح المعنى وسهولة الفهم والخلّو من اللبس، والتعبير بشكل سليم عن المعنى المراد دون انحراف أو إخلال به بدءً من الاختيار السليم لمفردات الجملة وانتهاءً بتجنب التراكيب التي قد تتسبب في اللبس الدلالي وهنا يبرُز التداخل بين التركيب والمعنى.
- جانب التركيب النحوي: والذي يهتم باختيار التركيب المناسب للجملة المولدة هل هو تركيب فعلي أم اسمي أم غير ذلك؟ إن كان فعلي فيجب حينئذٍ تحديد القالب الذي ستصاغ فيه الجملة المولّدة آليًّا إن كان في شكل (فعل-فاعل-مفعول) أم (فاعل-فعل-مفعول) وهكذا. وإن كان اسمي فيجب تحديد كيف يكون الترتيب بين الكلمات وبعضها داخل الجملة من تقديم لكلمة على أخرى أو تأخير كلمات بعينها وغير ذلك من العمليات النحوية من حذف وإضمار وغيرها.
- جانب الصرف: وهذا الجانب يعنى بكل كلمة داخل الجملة من الناحية المورفولوجية وتوليدها بما يتناسب مع سياق الجملة والكلمات المجاورة لها فيهتم بحالات المطابقة بين الفعل والفاعل والصفة والموصوف، وتصريف الأفعال والأسماء، والعلامات الإعرابية للكلمات طبقًا لموقعها داخل الجملة، وهنا يبرز التداخل بين التركيب والصرف. وغير ذلك من العمليات

المورفولوجية المختلفة. لذلك يُعدّ التوليد الآلي أحد المجالات المتقدمة في المعالجة الآلية للنصوص لما يشتمل عليه من عمليات مُعقدة تجمع بين التركيب النحوي والصياغة الدلالية للجملة والشكل المورفولوجي للكلمات.

وفي إطار لغة الشبكات الدلالية الحاسوبية العالمية تُستخدم أداة التوليد الآلي والتي يُطلق عليها (EUGENE[1]) لتوليد النصوص الطبيعية آليا من أية شبكة دلالية باستخدام موارد اللغة المراد توليدها من قواميس وقواعد توليد والتي تشتمل على ستة مراحل تبدأ بالمرحلة الأولى وهي تحديد الكلمة المناسبة لسياق الجملة تليها المرحلة الثانية وهي تعديل الشبكة الدلالية الناتجة عن عملية التحليل بما يتناسب مع اللغة الهدف ثم المرحلة الثالثة والتي يتم فيها استبدال العلاقات الدلالية بين الكلمات بعلاقات نحوية تعبر عن الدور النحوي لكل كلمة داخل الجملة لتمثيل البنية العميقة للجملة. ثم المرحلة الرابعة والتي يتم فيها استخلاص الشجرة النحوية السطحية من البنية العميقة للجملة تليها المرحلة الخامسة حيث يخضع هذا الشكل الشجري للعديد من عمليات التحويل والتغيير ليصبح في شكل قائمة أفقية من الكلمات. وأخيرًا المرحلة السادسة التي تعنى بتنقيح القائمة الأفقية لتوليد الكلمات في الشكل المورفولوجي المناسب للسياق طبقًا لقواعد كل لغة طبيعية لتتولد في النهاية الجملة الطبيعية التي كانت ممثلة في شكل شبكة دلالية.

- **التوليد الآلي للجملة العربية باستخدام لغة الشبكات الدلالية الحاسوبية العالمية**

فيما يلي عرض تفصيلي لمراحل التوليد الآلي لجملة عربية من الشبكة الدلالية الموجودة في الشكل (٣-٢٠) والتي تتكون من ثلاث علاقات دلالية تربط بين الفعل وفاعله الدلالي (agt)- الفعل ومفعوله الدلالي (obj) - الفعل وحاله (man).

١-هذا الرمز اختصار لـ (dEep-to-sUrface GENErator) ويُمكن استخدام أداة التوليد الآلي عن طريق هذا الرابط: http://dev.undlfoundation.org/generation/index.jsp.

```
{unl}
agt(201168468:0M.@present.@entry,110285313:00.@def)
obj(201168468:0M.@present.@entry,107739125:02.@def)
man(201168468:0M.@present.@entry,400105603:06)
{/unl}
```

الشَّكل ٣-٢٠: شبكة دلالية.

- **تحديد الكلمة المناسبة لسياق الجملة: تحويل المفاهيم إلى كلمات**

أولى مراحل التوليد الآلي هي مرحلة تحويل المفاهيم الموجودة داخل الشبكة الدلالية إلى كلمات عربية تناسب سياق ومعنى الجملة لتصبح الشبكة الدلالية بعد الانتهاء من تلك المرحلة كما نراها في الشكل (٣-٢١).

```
agt("0:"أكلM.@entry.@present, "00:"ولد.@def)
obj("0:"أكلM.@entry.@present, "02:"تفاحة.@def)
man("0:"أكلM.@entry.@present, "06:"بسرعة)
```

الشَّكل ٣-٢١: الشبكة الدلالية بعد تحويل المفاهيم العالمية إلى كلمات عربية.

- **تعديل الشبكة الدلالية بما يتناسب مع متطلبات توليد الجملة العربية**

تتيح قواعد التوليد الآلي لنظام لغة الشبكات الدلالية إمكانية تعديل الشبكة الدلالية الناتجة عن التحليل الدلالي بما يتناسب مع متطلبات كل لغة طبيعية لكنها مرحلة اختيارية قد نحتاج إليها وقد لا نحتاج إليها تبعًا لطبيعة الشبكة الدلالية الناتجة. وعدم المرور بتلك المرحلة لا يُعد إخلالاً بخطوات توليد الجملة. والشبكة الدلالية التي معنا ليست بحاجة إلى تعديل لذلك سيتم الاستغناء عن تلك المرحلة في هذا المثال.

- **تحويل العلاقات الدلالية إلى علاقات نحوية**

باستخدام قواعد تلك المرحلة يتم تحويل الشبكة الدلالية إلى شبكة نحوية باستبدال العلاقات الدلالية بين الكلمات بعلاقات نحوية. وبالنسبة للشبكة الدلالية الموجودة في الشكل (٣-٢١) تتحول علاقة الفاعل الدلالي (agt) إلى العلاقة النحوية مخصص فعلي ((Verb Specifier (VS)) وتتحول علاقة المفعول الدلالي (obj) إلى العلاقة النحوية متمم فعلي ((Verb Complement (VC))، وأخيرًا تتحول علاقة

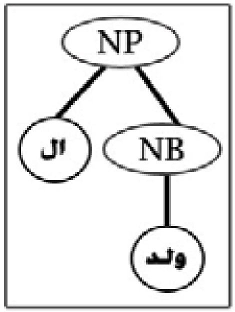
الحال الدلالي (man) إلى العلاقة النحوية ملحق فعلي ((Verb Adjunct (VA) أي حال الفعل. كما يظهر في الشكل (٣-٢٢).

VS("0:"أكلM.@entry.@present, "00:"ولد.@def)
VC("0:"أكلM.@entry.@present, "02:"تفاحة.@def)
VA("0:"أكلM.@entry.@present, "06:"بسرعة)

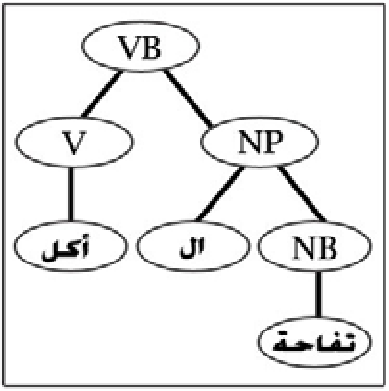
الشَّكل ٣-٢٢: الشبكة النحوية الناتجة عن المرحلة الثانية للتوليد.

- **بناء الشجرة النحوية: من البنية العميقة إلى البنية السطحية للجملة**

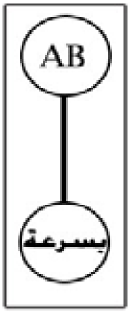
بعد تحويل العلاقات الدلالية (الشبكة الدلالية) إلى علاقات نحوية (الشبكة النحوية) يتم تمثيل تلك العلاقات النحوية في شكل شجري اعتمادا على نظرية X-Bar النحوية ولكي نصل إلى هذا الشكل الشجري لا بد من المرور بمراحل عدة في الانتقال بالجملة من البنية العميقة إلى البنية السطحية. فالعلاقات الموجودة داخل الشبكة النحوية الناتجة عن المرحلة السابقة عبارة عن علاقات ثنائية بين الكلمات ولكي نصل إلى الشكل الشجري لا بد من ربط تلك الأفرع الثنائية لتتكون الشجرة النحوية تدريجيًّا. ومن خلال الشبكة النحوية الناتجة عن المرحلة السابقة والموجودة في الشكل (٣-٢١) يتضح أن لدينا شبكة نحوية تتكون من مخصص فعلي (VS) ومتمم فعلي (VC) وملحق فعلي (VA). فنبدأ عن طريق العلاقة النحوية (VS) ببناء فرع المخصص الفعلي الذي هو عبارة عن الاسم «ولد» الموسوم بالسمة @def - أي أنه مُعرف بالألف واللام - وبالتالي يكون فرع المخصص الفعلي عبارة عن مركب اسمي نهائي (NP) كما في الشكل (٣-٢٣). ثم عن طريق العلاقة النحوية (VC) نبني فرع المركب الفعلي الوسيط (VB) هذا الفرع يجمع بين الفعل «أكل» ومتممه «التفاحة» - الذي هو أيضًا عبارة عن الاسم «تفاحة» موسومة بالسمة (@def) أي أن المتمم «تفاحة» عبارة عن مركب اسمي نهائي (NP) - ليصبح فرع المركب الفعلي الوسيط كما نراه في الشكل (٣-٢٤). وأخيرًا بناء فرع الملحق الفعلي ((A-Bar (AB) الذي هو عبارة عن الحال 0«بسرعة» كما في الشكل (٣-٢٥).



الشَّكل ٣-٢٣: المخصص الفعلي.

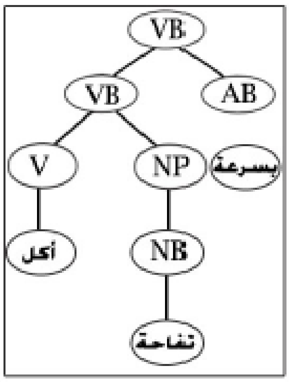


الشَّكل ٣-٢٤: المركب الفعلي الوسيط.

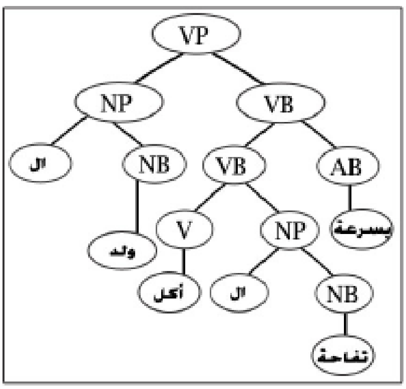


الشَّكل ٣-٢٥: الملحق الفعلي.

لازلنا في مرحلة بناء الشبكة النحوية؛ بعد بناء الأفرع بشكل منفرد تبدأ مرحلة جمع تلك الأفرع المنفردة لتكتمل الشجرة النحوية. فنجمع فرع المركّب الفعلي الوسيط (VB) مع فرع الملحق الفعلي (AB) وننتقل بهم لمستوى أعلى في الشجرة النحوية وهو مستوى المركب الفعلي الوسيط الثاني (VB) كما يظهر في الشكل (٣-٢٦). ثم نجمع فرع المركب الفعلي الوسيط الثاني (VB) الناتج عن الخطوة السابقة مع فرع المخصص الفعلي (NP) الموجود في الشكل (٣-٢٣) لنصل إلى رأس الشجرة النحوية التركيب الفعلي النهائي (VP) كما تظهر لنا في الشكل (٣-٢٧) وبذلك تكون قد اكتملت الشجرة النحوية وهي الهدف من هذه المرحلة.



الشَّكل ٣-٢٦: فرع المركب الفعلي الوسيط الثاني.



الشَّكل ٣-٢٧: الشجرة النحوية مكتملة.

- **فك الشجرة النحوية إلى تمثيل أفقي**

بعد اكتمال بناء الشجرة النحوية والانتهاء من وضع أسس البنية السطحية للجملة العربية تبدأ مرحلة فك الشجرة النحوية والانتقال بالجملة من التمثيل الشجري إلى التمثيل الأفقي ووضع الكلمات العربية بجانب بعضها البعض مع مراعاة المسافات فيما بينها حتى لا تخرج متلاصقة. وفي تلك المرحلة يجب علينا أن نأخذ في الاعتبار الترتيب بين كلمات الفرع الواحد بعد أن قمنا بالترتيب بين الأفرع في المرحلة السابقة. فالصفة تتبع الموصوف وأداة التعريف تسبق المعرف. وبعض الكلمات لا تحتاج إلى مسافة بينها وبين الكلمة التي تليها كأداة التعريف «ال» وغيرها. يوضح الشكل (٣-٢٨) التمثيل الأفقي للشجرة النحوية الناتجة عن المرحلة السابقة وهو عبارة عن جملة إسمية تبدأ بالفاعل «الولد» .

الولد أكل التفاحة بسرعة

الشَّكل ٣-٢٨: التمثيل الأفقي للشجرة النحوية.

- **معالجة الكلمة مورفولوجيًّا وتوليد الجملة في شكلها النهائي**

من أبرز الخصائص التي تطوع اللغة العربية للمعالجة الآلية طبيعتها الصرفية الاشتقاقية ذات الإنتاجية العالية والمستمدة أساسا من المكونين الرياضيين الجذر والوزن، ومن شأن هذا أن يدحض بعض الدعوات التي ترى أن اللغة العربية لغة معقدة وصعبة على مستوى المعالجة الآلية فرغم أنها تتسم بالاشتقاق الصرفي الغزير إلا أنه اشتقاق شبه منتظم. وهذا الانتظام يجعل المستوى الصرفي أكثر المستويات اللسانية قابلية للحوسبة دونًا عن المستويات الأخرى الدلالية والتركيبية خاصة في مرحلة التوليد الآلي للغة العربية. وفي المثال الذي لدينا الفعل «أكل» لديه السمة present@ والتي تدل على أن هذا الفعل حدث في زمن المضارع، ومع الأخذ في الاعتبار لطبيعة الفاعل المفرد المذكر تقوم قواعد الاشتقاق الخاصة بتلك المرحلة بتوليد شكل المضارعة للفعل «أكل» وهو «يأكل» ليصبح الشكل النهائي للجملة المولدة آليًّا كما في الشكل (٣-٢٩). كذلك تهتم تلك المرحلة بوضع اللمسات الأخيرة في توليد الجملة العربية والعناية بالشكل النهائي لها من خلال إضافة لعلامات استفهام أو أدوات تعجب أو تعديل كلمة ما بشكل معين لم نتمكن من تعديله خلال المراحل السابقة.

الولد يأكل التفاحة بسرعة

الشَّكل ٣-٢٩: الجملة العربية في شكلها النهائي.

٥- تطبيقات المعالجة الآلية للدلالة باستخدام لغة الشبكات الدلالية الحاسوبية العالمية

بعد أن استعرضنا كيف تقوم لغة الشبكات الدلالية الحاسوبية العالمية باستخلاص المعنى الدقيق للمحتوى الوارد في النصوص المكتوبة بأية لغة طبيعية وتمثيله في شكل حيادي مجرد، وكيف يستطيع النظام نفسه وضع هذا التمثيل الحيادي في إطار أية لغة طبيعية مرة أخرى، ينبغي علينا أن نتساءل كيف يمكن الاستفادة من مثل هذه التقنية. ولعل أول تطبيق يتبادر إلى أذهاننا هو الترجمة الآلية من لغة إلى أخري. وبالرغم من أن الترجمة الآلية تعد من أبرز استخدامات لغة الشبكات الدلالية الحاسوبية العالمية إلا أنها ليست التطبيق الأوحد بل بإمكان لغة الشبكات الدلالية الحاسوبية العالمية أن تنطلق لأبعد من هذا بكثير لأنها تقوم بفهم المعنى الذي تحمله الجمل والكلمات قبل تحويلها ونقله إلى لغة وسيطة نستطيع أن ننطلق منها إلى تطبيقات أخرى عديدة مثل البحث الآلي، والتلخيص الآلي، والتنقيح الآلي. وفيما يلي عرض لبعض من التطبيقات التي قامت على نظام لغة الشبكات الدلالية الحاسوبية العالمية.

٥,١- الترجمة الآلية للنصوص

نظرًا لما تحتله تطبيقات الترجمة الآلية للنصوص من أهمية في المعالجة الآلية للغات الطبيعية، ونظرا للجهد المبذول من اللغويين في المحاولة للوصول إلى تطبيق آلي يتمتع بإمكانية فهم النص واستيعابه ومن ثم التعبير عن معناه المقصود في الشكل النحوي الذي تسمح به اللغة الهدف، فإن الترجمة الآلية تُعد من أبرز التطبيقات التي يُمكن للغة الشبكات الدلالية الحاسوبية العالمية أن تساهم في تطويرها نظرًا لما تتمتع به من أدوات لغوية تمكنها من تحليل النص ونقل معناه المراد إلى لغة وسيطة وإعادة توليده ثانيةً في الشكل الذي يتماشى مع اللغة الهدف دون المساس بالمحتوى الدلالي للنص الأصلي. وللمركز العربي للغة الشبكات الدلالية الحاسوبية العالمية عدة محاولات في استخدام

لغة الشبكات الدلالية كمترجم آلي من أية لغة طبيعية إلى اللغة العربية نذكر منها على سبيل المثال:

- **موسوعة دعم نظم الحياة (EOLSS)**: هي موسوعة متعددة التخصصات وتعد أضخم موسوعة إلكترونية مكتوبة باللغة الإنجليزية حيث تتكون من ١٢٣٠٠٠ صفحة إنترنت أي ما يعادل حوالي ٢٥٠٠٠٠ صفحة مطبوعة. وقد قام المركز العربي للغة الشبكات الدلالية بمكتبة الإسكندرية بالاشتراك في المشروع الذي طُرح من قبل منظمة اليونسكو بهدف ترجمة هذه الموسوعة إلى اللغات الست الرسمية للأمم المتحدة، وكان المركز العربي مسئولا عن إنتاج النسخة العربية[١]. وتضمنت المرحلة الأولى من المشروع ترجمةَ ٢٥ نصا (حوالي ١٣٠٠٠ جملة) من الموسوعة مُشَفَّرًا من قبل مؤسسة لغة الشبكات الدلالية الحاسوبية العالمية (UNDL Foundation) ومُعَدًّا للترجمة إلى اللغات المطلوبة. وقد تمكنت قواعد التوليد العربية باستخدام قاموس متخصص من ترجمة النصوص المطلوبة. وتم تقييم الترجمات المولدة مقارنة ببعض أنظمة الترجمة الآلية الأخرى وكانت جودة النسخة العربية المولدة من لغة الشبكات الدلالية مرضية، وبالفعل تم إنشاء موقع على الإنترنت[٢] يضم النصوص المتفق على ترجمتها وتتصل فيه اللغات المشتركة في المشروع ببعضها، ويتم من خلال الموقع ترجمة النصوص على الإنترنت مباشرة. وبعد نشر نتائج هذا المشروع تلقى المركز العربي للغة الشبكات الدلالية دعوة من القائمين على بناء وتطوير موسوعة الحياة (EOL)[٣] لإصدار النسخة العربية من الموسوعة وهي موسوعة إلكترونية متاحة مجانا على الإنترنت باللغة الإنجليزية وتضم معلومات عن ١,٨ مليون كائن حي. وتهدف ترجمة الموسوعة إلى نشر المعرفة عن هذه الكائنات للمحافظة عليها، وقد تم اختبار مدى قدرة قواعد التوليد العربية على التعامل مع نصوص هذه الموسوعة، وكانت النتائج جيدة مقارنة ببعض أنظمة الترجمة الآلية الأخرى.

١- نتائج هذا المشروع موجودة في http://www.eolss.net/Eolss-Definition-Context.aspx.

٢- هذا الموقع هو http://www.undl.org/unleolss/unleolss.htm.

٣- لمعرفة المزيد عن موسوعة الحياة: http://eol.org.

كما قامت العديد من المراكز الأخرى الممثلة للغات المشاركة في مشروع لغة الشبكات الدلالية الحاسوبية العالمية ببناء أنظمة ترجمة آلية معتمدة على لغة الشبكات العالمية كما في المركز الهندي للغة الشبكات الدلالية، والمركز الروسي للغة الشبكات الدلالية، والمركز الفرنسي، وغيرهم.

٥,٢- الباحث الآلي عبر حاجز اللغة

استخدام آخر لا يقل أهمية عن توليد النصوص هو استخدام الشبكة الدلالية من أجل البحث داخل محتوى الإنترنت واسترجاع ما يحتاجه المستخدم من معلومات.

فعن طريق فهمه لما يبحث عنه المستخدم يستطيع البرنامج المبني على تكنولوجيا لغة الشبكات الدلالية الحاسوبية العالمية البحث عن المعلومات المطلوبة داخل صفحات الإنترنت المكتوبة بأي لغة وليست لغة البحث فقط بينما يقوم بإظهار نتائج البحث بلغة المستخدم للإنترنت أيا كانت اللغة الأصلية المخزنة بها تلك النتائج في صفحات الإنترنت.

أي أن برنامج لغة الشبكات الدلالية سيعتمد على مقارنة المعنى الدلالي. وكذلك يعتبر استرجاع المعلومات عبر اللغة (cross-language information retrieval) من ضمن التطبيقات التي تندرج تحت البحث عبر اللغة. حيث يمكن على سبيل المثال البحث داخل فهارس المكتبات الإلكترونية ومعرفة معلومات عن الكتب المتاحة والحصول على أي معلومة من هذه الكتب بل وقراءتها مهما كانت لغة هذه الكتب.

ولقد قام المركز العربي بتصميم وتنفيذ نموذج لهذا الباحث الآلي على تطبيق «نظام مكتبات غير معتمد على لغة»، وهو نظام يسمح بترجمة المعلومات الخاصة بالكتب إلى اللغات الست الرسمية للأمم المتحدة بالإضافة إلى اللغة البرتغالية.

وهذا النظام مصمم لكي يسمح للمستخدم باستدعاء وتصفح المعلومات الخاصة بالكتب الموجودة في فهارس المكتبات الإلكترونية باللغة التي يطلبها بصرف النظر عن اللغة المخزنة بها كما يسمح للمستخدم المتخصص (المكتبي) بفهرسة الكتب وإضافة أو تعديل المعلومات المختلفة الخاصة بكل كتاب، كما يوفر معلومات إحصائية عن عدد الكتب التي تم تخزينها.

كما قام المركز الفرنسي للغة الشبكات الدلالية الحاسوبية العالمية بعمل نظام بحث واسترجاع للمعلومات عبر الإنترنت معتمدا على لغة الشبكات الدلالية الحاسوبية العالمية. وكذلك المركز الإسباني للغة الشبكات الدلالية الحاسوبية العالمية الذي وضع نظامًا متعدد اللغات لاسترجاع للمعلومات [٣٠] وكذلك أيضا المركز الهندي للغة الشبكات الدلالية الحاسوبية العالمية [٣٢].

٥,٣- التلخيص والتنقيح الآلي للنصوص

لا تقتصر استخدامات لغة الشبكات الدلالية الحاسوبية العالمية على التحويل من لغة إلى أخرى، فقد تستخدم في داخل إطار اللغة الواحدة. وفي تلك الحالة يكون التحويل من شكل إلى شكل أو من أسلوب إلى أسلوب لكن بنفس اللغة ودون المساس بالمحتوى.

من بين الاستخدامات التي تندرج في هذا الإطار: التنقيح الآلي، التلخيص الآلي، والتبسيط الآلي. وفي عمليات التنقيح الآلي يتم تغيير بعض المفردات أو بعض التراكيب في النص الأصلي للوصول إلى نسخة مختلفة معدلة، فمثلا إبدال بعض المفردات العامية بأخرى فصحى أو العكس فيكون النص الخارج نصا مختلفا من المنظور الاجتماعي أو التنوع المفردي.

وبنفس الطريقة يمكن تغيير الأسلوب الأدبي العام عن طريق تبسيط بعض التراكيب أو إضافة بعض المحسنات البلاغية التي لا تغير في المحتوى الدلالي للنص وبذلك يتولد نص مختلف عن النص الأصلي من حيث الطابع الأدبي.

ويمكن أيضا استخدام الفهم الذي تصل إليه لغة الشبكات الدلالية الحاسوبية العالمية في عملية التلخيص الآلي عن طريق توليد نص مقابل يختلف مع النص الأصلي في طوله بحيث يكون مختصرا. ويتم ذلك من خلال تحديد المفاهيم الرئيسية والمفاهيم الثانوية والاستغناء عن تلك الثانوية.

أما مهمة التبسيط الآلي فهدفها جعل النص الأصلي أسهل من ناحية القراءة والفهم. ويحدث هذا من خلال تغيير بعض الرموز أو التراكيب التي من شأنها تعقيد النص.

فمثلا في الجملة التالية: «ولد في بهجورة - الأقصر - مصر» تكمن الصعوبة في تحديد مدلول العلامة «-» ، عندئذ تكون مهمة لغة الشبكات الدلالية الحاسوبية العالمية معرفة ما المقصود بهذه العلامة والتعبير عنها بشكل أكثر وضوحا فيكون العنوان المبسط «ولد في بهجورة في الأقصر في مصر» .

وهذا التفسير لا يعتمد على العلامة نفسها بل يعتمد على فهم لغة الشبكات الدلالية الحاسوبية العالمية لمعنى الجملة، فنفس تلك العلامة في جملة مثل «معارض كثيرة بفرنسا-كندا» يتم تبسيطها إلى «معارض كثيرة بفرنسا وكندا» .

وتقوم بكل العمليات السابقة أداة موجودة بالفعل لكنها لازالت تخضع للتطوير وهي أداة توت[1]؛ وهي عبارة عن مكتبة رقمية للنصوص الممثلة في شبكات دلالية، وتضم أكثر من ٣٠٠٠٠ عنوان والشبكات الدلالية الممثلة لهم (إن وجدت).

وبإمكان المستخدم اختيار عرض أي من النسخة الأصلية، أو النسخة المختصرة، أو النسخة المنقحة، أو النسخة المبسطة.

٦- دعوة للمشاركة

من أجل تطوير المعالجة الآلية للدلالة في اللغة العربية قمنا بوضع خطة طريق للباحثين في هذا المجال بهدف تحقيق أفضل النتائج في معالجة الدلالة. تتضمن تلك الخطة العديد من النقاط البحثية، منها:

١- توصيف الأدوار النحوية في الجملة العربية بما يقابلها من أدوار دلالية وكيفية الانتقال من الدور النحوي إلى الدور الدلالي والعكس.

٢- دراسة المتعلقات النحوية والتصنيف الدلالي للكلمات العربية أيًّا كانت فئتها المعجمية (الأسماء، الأفعال، الصفات، الظروف).

٣- دراسة كيفية تخزين الكلمات المركبة داخل القاموس واشتقاق الأشكال الصرفية المختلفة منها.

١- لمعرفة المزيد عن الأداة «توت» يرجى اتباع هذا الرابط: http://www.unlweb.net/tut.

٤- بناء شبكة الكلمات العربية الدلالية (Arabic WordNet).

٥- دراسة أسس ومعايير بناء قاموس حاسوبي عربي، يضم كلمات اللغة العربية وما تحتاجه من معلومات لُغَوِيَّة؛ يصلح لتطبيقات المعالجة الآلية للغة العربية.

٦- دراسة عن كيفية التعامل مع المركبات اللفظية التي يفصل السياق بين أجزاءها مما يؤدي إلى تباعدها وبالتالي صعوبة التعرف عليها. مثل تعبير «قطع مسافة» والذي قد يأتي بهذا الشكل: (قطعت السيارة مسافة ميلين قبل أن تصل لوجهتها) أثناء التحليل الآلي لهذه الجملة سيكون هناك مشكلة في جعل الحاسوب يعتبر هاتان الكلمتان المنفصلتان مفهوم واحد.

٧- دراسة التراكيب الدلالية التي لا تقبل التجاور وشروط التجاور. فهناك بعض التراكيب الدلالية التي يصعب تتابعها داخل الجملة العربية مثل تتابع الصفة بعد الفعل غير مسموح به داخل الجملة العربية كما في المثال: (ذهبت الجميلات إلى الحديقة) الجميلات هنا اسم وليست صف فيكون على المحلل الآلي أن يختار «الجميلات» ذات وسم الاسم وليس الصفة.

ببليوجرافيا مرجعيَّة

١. الأنصاري (سامح)، ناجي (مجدي)، العدلي (نهى): النظام العربي للغة التواصل العالمية، المؤتمر الدولي لعلوم وهندسة الحاسوب (ICCA)، الرياض، المملكة العربية السعودية، ٢٠١١.

٢. أيُّوب (عبد الرَّحمن): التحليل الدلالي للجملة العربية، المجلة العربية للعلوم الإنسانية، جامعة الكويت، مارس ١٩٨٣.

٣. الجرجاني (عبد القاهر): دلائل الإعجاز في علم المعاني، دار الكتب العلمية، بيروت، لبنان، ط١، ١٩٨٨.

٤. كريستوفر س . بتلر: اللغة والحسابية، ضمن الموسوعة اللغوية، تحرير: د. ن. ي. كولنج، ترجمة د. محي الدين حميدي ود. عبدالله الحميدان، المجلد الثاني.

5. Adly, N. & Sameh Alansary, S. (2009c). Evaluation of Arabic Machine Translation System based on the Universal Networking Language, 14th International Conference on Applications of Natural Language to Information Systems (NLDB 2009), Saarland University, Saarbrücken-Germany.

6. Alansary, S. & Nagi, M. & Adly, N. (2006). Processing Arabic Text Content: The Encoding Component in an Interlingual System for Man-Machine Communication in Natural Language, 6th International Conference on Language Engineering, Ain Shams University, Cairo.

7. Alansary, S. & Nagi, M. & Adly, N. (2007). A Semantic-Based Approach for Multilingual Translation of Massive Documents, 7th Symposium of Natural Language Processing, Thailand.

8. Alansary, S. & Nagi, M. & Adly, N. (2009a). A Library Information System (LIS) based on UNL knowledge infrastructure, Seventh International Conference on Computer Science and Information Technologies, Yerevan, Armenia.

9. Alansary, S. & Nagi, M. & Adly, N. (2009b). The Universal Networking Language in Action in English-Arabic Machine Translation, 9th Conference on Language Engineering, Ain Shams University, Cairo.

10. Alansary, S. & Nagi, M. & Adly, N. (2010a). UNL+3: The Gateway to a Fully Operational UNL System, 10th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt.

11. Alansary, S. & Nagi, M. & Adly, N. (2011). Understanding Natural Language through the UNL Grammar Work-bench, Conference on Human Language Technology for Development (HLTD 2011), Bibliotheca Alexandrina, Alexandria, Egypt.

12. Alansary, S. & Nagi, M. & Adly, N. (2011). UNL Editor: An Annotation tool for Semantic Analysis, 11th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt.

13. Alansary, S. & Nagi, M. & Adly, N. (2012). IAN: An Automatic Tool for Natural Language Analysis, 11th International Conference on Language Engineering , Ain Shams University, Cairo, Egypt.

14. Alansary, S. & Nagi, M. & Adly, N. (2013). A Suite of Tools for Arabic Natural Language Processing: A UNL Approach, (ICCSPA'13), Sharjah, UAE.

15. Alansary, S. & Nagi, M. (2013). LILY: Lan-guage-to-Interlanguage-to-Language System Based on UNL, 12th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt.

16. Alansary, S. (2009). Issues on Interlingua Machine Translation Systems, 9th Conference on Language Engineering, Ain Shams University, Cairo, Egypt.

17. Alansary, S. (2010b). A Practical Application of the UNL+3 Program on the Arabic Language, 10th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt.

18. Alansary, S. (2012). A Formalized Reference Grammar for UNL-based Machine Translation between English and Arabic, COLING, the 24th International Conference on Computational Linguistics, IIT Bombay, Mumbai, India.

19. Alansary, S. (2014). MUHIT: A Multilingual Harmonized Dictionary, The 9th edition of the Language Resources and Evaluation Conference, 26-31 May, Reykjavik, Iceland.

20. Alansary, S. (2015). Keys: A Knowledge Extraction System Based on UNL knowledge infrastructure, TENCON - IEEE Region 10 Conference, Macau, China.

21. Bhat, B. & Bhattacharyya, P. (2011). IndoWordnet and its Linking with Ontology, International Conference on Natural Language Processing (ICON 2011), Chennai.

22. Boguslavsky, I & Frid, N. & Iomdin, L. & Kreidlin, L. & Sagalova, I. & Sizov, V. (2000). Creating a Universal Networking Language Module within an Advanced NLP System, COLING 2000, Saarbrücken, Germany 31/07-04/08, p.76-82.

23. Boitet, C. (2002). A rationale for using UNL as an Interlingua and more in various domains, proceedings “First International Workshop on UNL, other Interlinguas and their Applications, LREC2002, Las Palmas, Spain.

24. Boitet, C. (2005). Gradable quality translations through mutualisation of human translation and revision, and UNL-based MT and coedition.Universal Network Language: Advances in Theory and Applica-tions. Research on Computing Science 12, 2005, pp. 395–412.

25. Boudh, S. & Bhattacharyya, P. (2010). Unification of Universal Word Dictionaries Using WordNet Ontology and Similarity Measures, 5th International Conference on Global Wordnet (GWC2010), Mumbai.

26. Burton, R.R. & Woods, W.A. (1976). A compiling system for augmented transition networks. Preprints of COLING 76: The International Conference on Computational Linguistics, Ottawa.

27. Butler, C. S. (1985a). Computers in Linguistics, Blackwell, Oxford.

28. Butler, C. S. (1985b). Statistics in Linguistics, Blackwell, Oxford.

29. Butler, C. S. (1985c). Systemic Linguistics: Theory and Applications, Batsford.

30. Cardeñosa, J. & Gallardo, C. & Toni, A. (2009). Multilingual Cross Language Information Retrieval: A new approach, UNL Workshop, in conjunction with CSIT 7th International Conference Yerevan, Armenia.

31. Dave, S. & Parikh, J. & Bhattacharyya, P. (2001). Interlingua-based English- Hindi Machine Translation and Language Divergence. Journal of Machine Translation 16(4), 251–304 (2001) 8. Chatterji, S., Roy, D., Sarkar, S.

32. Kagathara, S. & Deolalkar, M. & Bhattacharyya, P. (2005). A Multi Stage Fall-back Search Strategy for Cross-Lingual Information Retrieval, proceedings of Symposium on Indian Morphology, Phonology and Language Engineering [SIMPLE 2005], IIT Kharagpur, India.

33. Martins, R. & Avetisyan, V. (2009). Generative and Enumerative Lexi-cons in the UNL Framework, In proceedings of 7th International Conference on Computer Science and Information Technologies, CSIT 2009, Yerevan, Armenia, 2009.

34. Schank, R. C. (1972). 'Conceptual dependency: a theory of natural lan-guage understanding', Cognitive Psychology, 3 / 4: 552 – 630.

35. Shaalan, K. & Rafea, A. & Baraka, H. (2006). Mapping Interlingua Representations to Feature Structures of Arabic Sentences, The Challenge of Arabic for NLP/MT, London.

36. Uchida, H. & Zhu, M. & Senta, T. D. (2005). Universal Networking Language, UNDL Foundation.

37. Uchida, H. (1996). UNL: Universal Networking Language – An Electronic Language for Communication, Understanding, and Collaboration, UNU/IAS/UNL Center, Tokyo, Japan.

38. Wilks, Y. 'Preference semantics', in Keenan, E. L. (ed.), Formal Se-mantics of Natural Language, Cambridge University Press, Cambridge: 329– 48,1975.

الفصل الرابع

موارد التَّعلُّم الآليّ

(مدخل إلى التَّعلُّم الآليّ)

د. مُحسِن رَشوان

١- شجرة القرار.

٢- مصنِّف بايز المبسط.

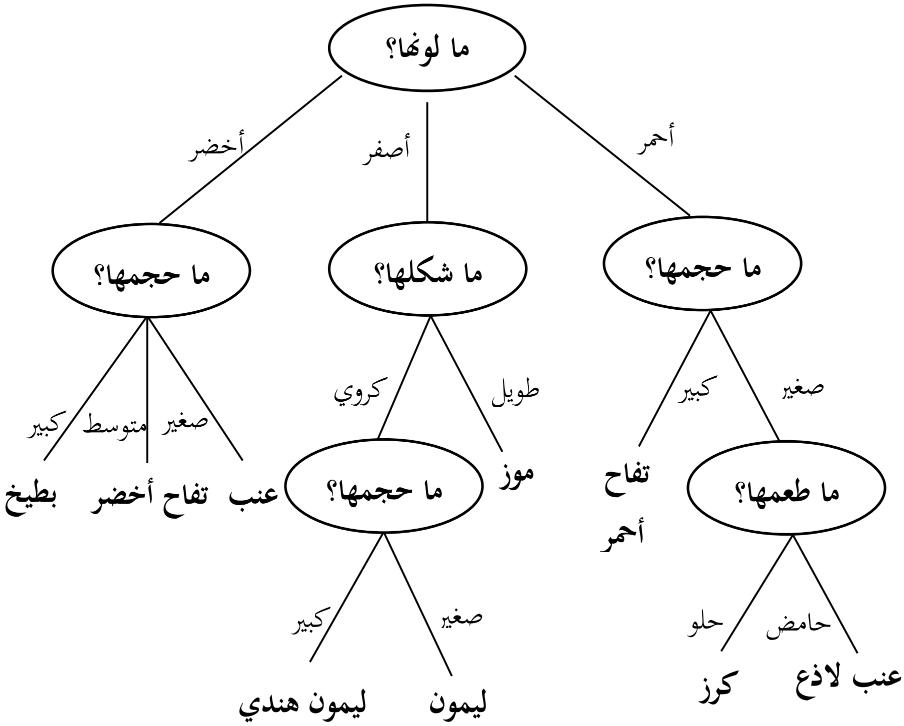
٣- الشبكات العصبية.

٤- آليَّات المتجهات الداعمة.

٥- نماذج ماركوف المُخَبَّأة.

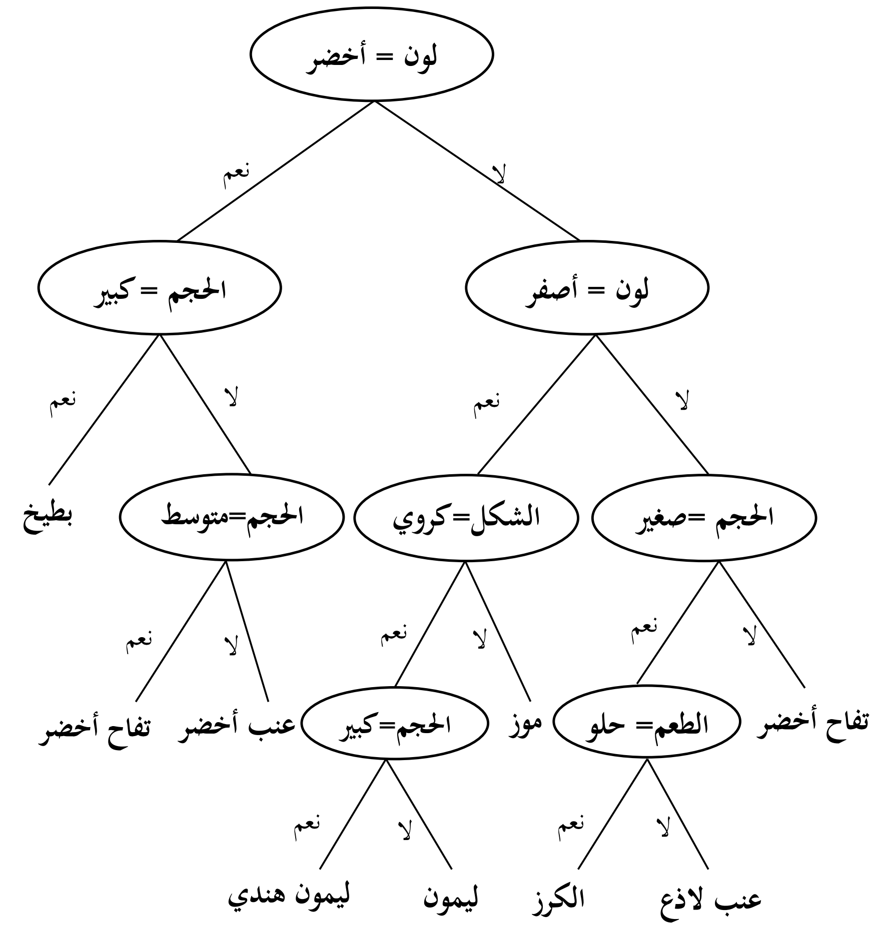
١ - شجرة القرار

في كثير من الأحيان يمكن حسم القرار في ميادين حوسبة اللُّغة عن طريق ما يسمى «شجرة القرار» (Decisions Tree). وكثيرًا ما تُستخدم شجرة القرار لحل المشكلات الَّتي تسمح طبيعتها بذلك؛ وإليك هذا المثال. لعبة «ما هي الفاكهة؟» سيُسمَحُ فيها بثلاثة أسئلة - بحدٍّ أقصى - للوُصُول إلى نوع الفاكهة. انظر الشكل (٤-١).



الشَّكل ٤-١: نوع الفاكهة.

وفي بعض الأحيان تُستَخدَمُ أسئلة بسيطة (من نوع: «نعم/ لا» فقط). ولتحويل الشجرة إلى هذا النوع البسيط من الأسئلة يمكن مراجعة نفس المثال السابق في صورته الجديدة في الشكل (٤-٢).



الشَّكل ٤-٢: شجرة القرار مبنية على نوع الأسئلة «نعم/ لا» - قد نحتاج إلى أكثر من ٣ أسئلة.

وتُستخدم شجرةُ القرار بكفاءةٍ مع حلولٍ تعتمد على القواعد. وفي كثير من الأحيان تحتاج هذه القواعد إلى تنظيم وترتيب، ويكون ذلك باستدعاء شجرة القرار.

٢- مصنِّف بايز المبسط

تعالوا معًا نصيغ المشكلة رياضيا، المطلوب هو حساب $P(s_i/C)$

حيث s_i هو المعنى i (Sense i) الذى يمكن أن تأخذه الكلمة w محل التحليل. ويقرأ التعبير السابق كالآتى: احتمال المعنى i بشرط توفر السياق C. وليس لدينا معلومات

يمكن أن نستنبط منها الحل مباشرة لـ $P(s_i/C)$ ولكن «بيز» يقدم لنا تحليلا مهما لابد أن نلجأ إليه كخطوة نحو الحل كما في المعادلة ١:

$$(1) \qquad P(s_i/C) = \frac{P(C/s_i)\, P(s_i)}{P\,(C)}$$

وهذا يقربنا خطوة نحو الحل، فحساب $P(C/s_i)$ ، $P(s_i)$ ممكن وسهل، اما بالنسبة $P(C)$ فلسنا في الحقيقة في حاجة إليها اصلاً لأنها ستكون موجودة مع كل المعانى المحتملة للكلمة w محل الدراسة. ولذلك فالمعادلة السابقة يعاد صيغتها كالآتى:

$$(2) \qquad g(s_i/C) = P(C/s_i)\, P(s_i)$$

تلاحظ هنا أننا غيرنا اسم الطرف الايسر إلى $g(s_i/C)$ لانه لم يعد يعبر عن الاحتمال بالمعنى المصطلحى الذى قيمته محصورة بين الصفر والواحد.

وهنا سنحتاج لفرض آخر لتبسيط الحل من خلال مصنَّفٍ بسيط، إلاَّ أنَّه فعَّالٌ لدرجة كبيرة، وتتنافس نتائجه -في كثيرٍ من الأحيان- مع نتائج مصنَّفاتٍ أخرى أكثر منه تعقيدًا. إنَّه «مصنف بايز المبسط (Naïve Bayes Classifier). وجريا على عرف الكُتَّاب عند استخدام مصطلح كثير الاستخدام أن يختصروا اسمه باستخدام الأحرف الأولى، أي (م ب م) ويختصرونه بالإنجليزية أيضًا (NBC). ويسمى المبسط لأن هناك فرضية رياضية لتبسيط الحل وهي اعتبار أن الكلمات التي تمثل السياق مستقلة بعضها عن بعض- وإن كان ذلك في الحقيقة غير صحيح، لأن بعض الكلمات يقترن كثيرًا بكلمات أخرى. وهذا الفرض سمح لنا بإمكانية التعامل مع السياق بشكل مبسط. والسياق هو مجموع الكلمات التي سبقت الكلمة مباشرة أو تلتها. ويجوز لنا بهذا الفرض أن نكتب سياق الكلمة w_j كالآتي:

$$(3) \quad P(C) = P(w_1) * P(w_2)\ldots P(w_{j-1}))*P(w_{j+1})\ldots P(w_N)$$

وكذلك يمكن إعادة كتابة المعادلة (٣) كالآتي:

$$(4) \quad g(sj/C) = [P(w_1/s_j) * P(w_2/s_j)\ldots P(w_{j-1})/s_j) \\ *P(w_{j+1})/s_j)\ldots P(w_N)]*P(s_j)$$

إن صياغة المعادلة يجعل الحل في متناول أيدينا. فلو أننا تمكنا من حساب الكميات ($P(w_k/s_j)$)، ثم حسبنا أيضًا $P(s_j)$ نكون قد حسمنا الأمر كله وعرفنا أي الحلول في هذا السياق هو الأوفق.

إن حساب هذه الكميات يمكن الرجوع إليه في ملحق-١ لنظرية الاحتمالات وكذلك فصل «نمذجة اللغة». ولايفوتنا هنا أن نذكر بأن الاحتمال $P(s_j)$ يسمى النحو الأحادي، وهو احتمال أن تأتي الكلمة بهذا الحل عموما، بصرف النظر عن السياقات المختلفة (أي: احتمال وجودها ككلمة مفردة).

ومثال ذلك: كلمة «قال» من مادة القول قد يصل نحوها الأُحادي - المشروط بورود «قال» - إلى أكثر من ٩٩٩, ٠ بينما كلمة «قال» من مادة قيل (أي النوم بالظهيرة) قد لا يصل نحوها الأحادي - المشروط بورود «قال» - إلى ٠٠١, ٠. والجدير بالذكر أننا سوف نقابل عند تطبيق هذا الخوارزم أو هذا المصنف مشكلة، وهي أن بعض الكلمات لم نرها من قبل في الذخيرة اللغوية الَّتي تدرب النظام عليها. وفي سياق جديد إذا أتت كلمة واحدة لم تُر من قبل، فسيكون احتمال ورودها صفرًا، وسوف نضرب في صفر فتكون النتيجة صفرا مهما كانت قوة شواهد الكلمات الأخرى في السياق. ولقد واجهنا هذه المشكلة في الباب الثَّامن - عندَ حَدِيثنا عن «نمذَجة اللُّغة» واستطعنا أن نمنع هذا الصفر بافتراض نسبة احتمال صغيرة نسبيا لما لم نره من الكلمات.

٣- الشبكات العصبية

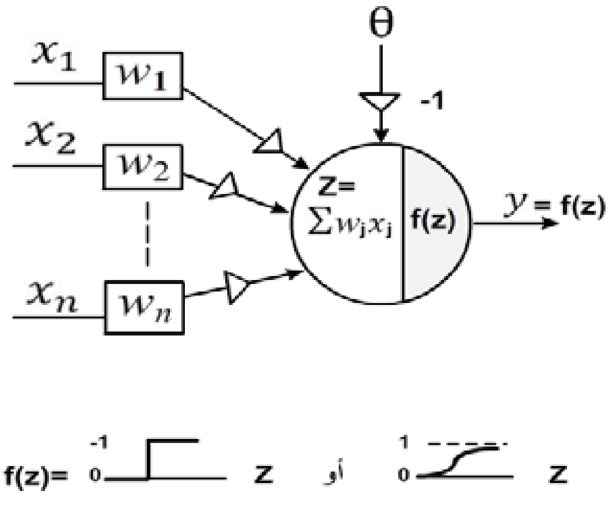
يتمتع الإنسان بمقدرة رائعة على البيان بمُختلف أنواعه، سمعا وبصرا وفهما ونطقا... إلخ. ولقد لَفَتَ الدماغُ البشريُّ أنظارَ العلماء، ولا يزالون مبهورين بقدرة المخ البشريّ على الإدراك والتَّذكُّر والاستماع والإبصار، فلم يحيطوا علمًا إلا بالقليل جدًّا عن كيفية ذلك. ولقد حاول بعض العلماء فهمَ الوحدة الأساسية واللبنة الأولى في بناء مخ الإنسان، ونقصد «الخلية العصبية».

يحتوي مخ الإنسان على نحو ١٠ مليارات خلية عصبية ولكل خلية منها اتصال بغيرها من الخلايا بمتوسطٍ يصل إلى نحو ١٠ آلاف وصلة بـ ١٠ آلاف خلية أخرى. وهذا يعني أن عدد وصلات الخلايا يبلغ نحو مئة ألف مليار وصلة. ويُظنُّ أن في هذه الوصلات تخزِّن المعلومات. وهذه الوصلات لا تبقى بلا استخدام، وإنَّما تتآكل وتضمر

مع الوقت وخاصة التي تبقي بلا استخدام. فعددها كامل عند الصغار وتقل مع الزمن. فالذي يحفظ القرآن صغيرًا عنده الكثير منها ليخزن فيه ما يحفظ، والذي يحفظ على الكبر يجد صعوبةً أكبر سواء في الحفظ أو تذكر ما يحفظ.

وبالنسبة للغة – وهي متطورة جدًّا عند الإنسان – فقد رُويَت حادثة عن طفلة في السادسة من عمرها، وقد حبسها أبوها وهي صغيرة جدا في قبوٍ تحت المنزل، وكان يلقي إليها الطعام دون أي مخالطة أو محادثة حتى بلغت السادسة من عمرها. وبعد اكتشاف هذه البنت (بالطبع عوقب أبوها)، أخذ علماء كثيرون البنت لينظروا – نفسيًّا ولغويًّا – ماذا فقدت؟ وكيف يمكن تعويضها؟. وحاولوا تعليمها اللغة شهورا طويلة فاستطاعت أن تستوعب أسماء الأشياء، مثل: شجرة، طريق، ثلاجة،... إلخ. ولكن تعبيرات مثل «في الثلاجة»، «إلى المدرسة»،... إلخ، لم تستطع تعلمها؛ فاستنبطوا أن الإنسان مزود بأداة للغة (جزء من الدماغ مخصص لذلك)، ولها وقتها للتعلم. فإذا مر الوقت المناسب ضعفت وتآكلت. ولعلها تلك الوصلات التي تتآكل إذا مر وقت استعمالها ولم تستعمل. لذلك من المهم جدا أن نعطي الأولاد حقهم في التعلم واللعب، ولكل سنٍّ ما يناسبه من الألعاب وما يناسبه من المفردات وقواعد للغة الَّتي يُلتَمَس تعلمها.

ولقد اجتهد العلماءُ ووضعوا نموذجا رياضيًّا مبسَّطا لعمل الخلية العصبية، كما تعرَّضوا لكيفيَّة الجمع بين طبقات الخلايا العصبيَّة، على النَّحو المبَيَّن في الشكل (٤–٣).



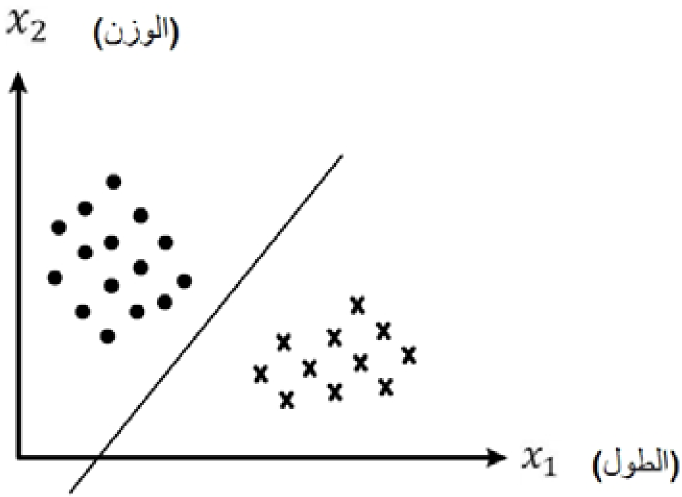
الشَّكل ٤–٣: النموذج الرياضيّ المبسَّط للخليَّة العصبيَّة.

إن الوصلات بين الخلايا (الشكل ٤-٣) تحمل الأوزان «w_j» لكل إشارة x_j داخلة للخلية؛ وإذا زاد المجموع المرجح (weighted sum) في حالتنا $\Sigma_{j=1}^{N} w_j x_j$ على ما يسمى «العتبة» (threshold) ويرمز لها بالرمز θ (وهي قيمة تتعلمها الخلية كما تتعلم الوصلات قيم الأوزان)، فإن الخلية تعطي خرجا «y» له قيمة عالية يُعبر عنها رياضيا بالقيمة «١»؛ وإلا فإن قيمة y تظل «٠».

والآن يمكننا أن نتعلم كيف تعمل الخلية العصبية للتمييز بين شكلين مثلا. فلو كانت لكل شكل ميزات مختلفة (الطول والعرض مثلا) فإننا نقيس هذه الميزات أو الخصائص ونضعها في متجه (Vector) من الخصائص:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

ثم نعيد هذه القياسات مرات عديدة لكل شكل على حدة. وتعالوا نفرض أن لدينا - من هذه الخصائص التي تقاس - اثنتين فقط (ليسهل التصور). لو تصورنا أن لدينا أولادًا وبنات في سن معينة، وكنا نقيس الطول والوزن ونحاول من خلالهما معرفة جنس الطفل مثلاً، فستكون قياسات الأولاد والبنات على النَّحو المبَيَّن في الشكل (٤-٤).

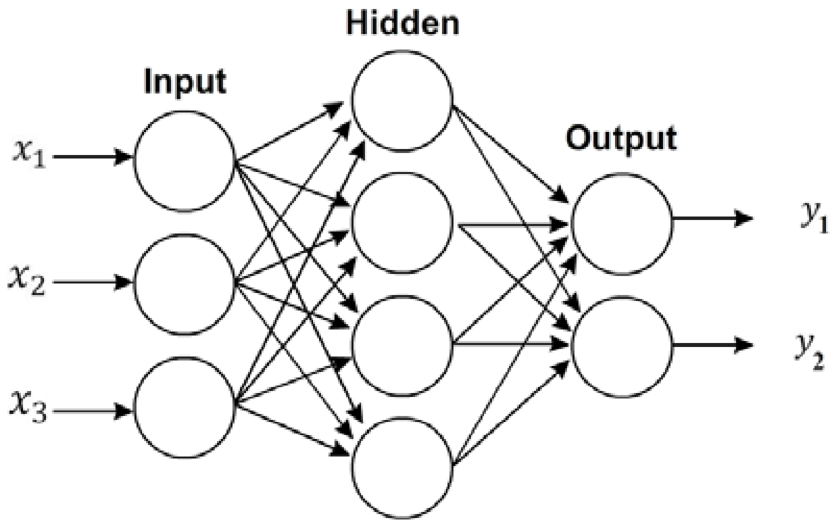


الشَّكل ٤-٤: عينات من الأولاد والبنات بعد قياس الطول والوزن لكل منهم (حيث «x» تمثل عينة من الأولاد و «•» تمثل عينة من البنات).

والمطلوب من النموذج الرياضي للخلية العصبية التفرقة بين الأولاد والبنات الذين أخذت قياساتهم، كما في الشكل (٤-٥). (ملاحظة: افتراض تمايز الجنسين بهذا الشكل تَخَيُّلِيّ بعض الشيء لغرض الشرح فقط).

فبتدريب النموذج الرياضي للخلية العصبية، فإنَّ قِيَمًا لـ w_1، w_2، θ يمكن الوصول إليها بالتدريب والتعلم حتى تفصل بين عينات الجنسين؛ فإذا كانت قيم (x_1، x_2) تمثل أولادًا، فإن قيمة الخرج «y» للخلية يكون «١»، وإلا فإنه يكون «٠». إذن، كيف يتمُّ تدريب الخلية؟ إن هذا أمر بسيط في الواقع، ويشبه تعليم الأطفال. إننا حين نعلم الأطفال نريهم الشكل ونقول هذا «كذا»، ونعيد ونكرر حتى يستطيع الطفلُ تمييز هذا الشكل وحده. نقوم بعملية مماثلة رياضيًّا حتى نتمكن من الوصول بالأوزان (θ،w_2، w_1) لتمثيل فاصل بين عينات الأولاد والبنات كما في الشكل (٤-٤).

حتى الآن يمكن - من خلال نموذج رياضي لخلية عصبية واحدة - التمييز بين شكلين بسهولة؛ ولكن حتى نتمكن من التمييز بين أشكال معقدة يجدر بنا أن نستخدم تراكيب معقدة وفي شكل طبقات للخلايا العصبية. انظر الشكل (٤-٥).



الشَّكل ٤-٥: الخلايا العصبية في شكل طبقات.

هذه الأشكال المركبة قادرة على تعلم التفرقة بين أشكال معقدة (وأكثر من شكلين في آن واحد). وعند تدريبها تستخدم طرق رياضية لتعليم الأوزان (weights) من خلال استخدام الخصائص المختلفة للأشكال المطلوب التعرف عليها. وتخيل عند كل سهم وزن «w» قابل للتعليم.

وتتميز الشبكات العصبية (Neural Networks) بخصائص جذابة للعاملين في حقل التمييز بين الأنماط، من أهمها:

١-أنَّ أعباء الحسابات تتوزع على كمية كبيرة من الخلايا العصبية، وكلها تعمل على التوازي فلا يعطل بعضها بعضا. وهذا مناسب للتطور الحادث في تقنيات الحواسيب، إذ إن هذه التقنية تتجه إلى استخدام كمية كبيرة من المعالجات (processors) التي يمكن استخدامها على التوازي.

٢-أن الأوزان (weights) التي تتعلمها تتوزع فيها المعلومة الواحدة على أوزان كثيرة؛ والدليل على ذلك أننا لو عطلنا (في الشكل ٤-٥) عددًا من نماذج الخلايا العصبية (مثلا ١٠٪ من المتاح منها - بغرض التجربة)، فغالبا ستظل تعمل بكفاءة تامة؛ وهذا بالضبط ما يحدث في مخ الإنسان، إذ تموت كل يوم خلايا ويظل المخ يعمل بكفاءة تامة، إلا إذا تأخر العمر ومات كثيرٌ جدًّا من هذه الخلايا، أو عند حدوث حادث يصيب خلايا المخ بشدة؛ عندئذ ربما تضعف هذه الكفاءة. هذه الخاصية مهمة جدًّا للكائنات الحية، لأنها تتعرض للإصابة والمرض مما يعطى فرصة لفقد بعض الخلايا، أو حتى لعامل الزمن. بينما في الحاسبات المألوفة لدى البشر لا تتحمل البرامج التقليدية أن تُفقدها أيَّ شيء، وإلا تعطلت عن العمل فورًا.

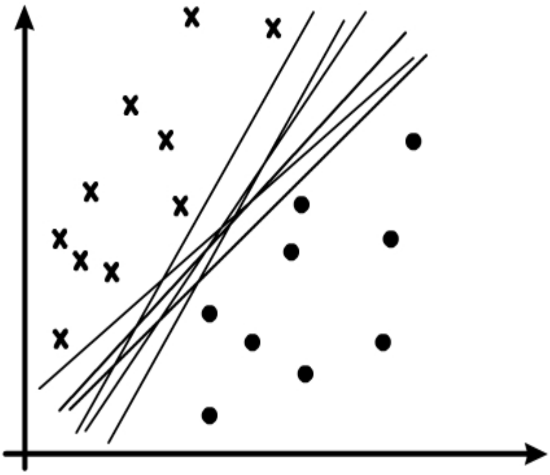
إلا أن هناك مشكلة تواجه الباحثين في مجال الخلايا العصبية، وهي أنهم عند تركيب عدد كبير منها لحل مشكلة بعينها لا يمكنهم الوصول للحل الأمثل، وإنما يحاولون الوصول إلى أحسن حل ممكن، وليس هناك ما يضمن أنه الحل الأمثل.

وهناك مشكلة أخرى، تكمن في أنهم لا يعرفون سلفًا طريقةً لتركيب هذه الخلايا حتى نضمن أحسن حل للمشكلة المراد استخدام النموذج الرياضي للخلايا العصبية

في حلها. أي، لا يعرفون عدد طبقات الخلايا وعدد الخلايا في كل طبقة - كل ذلك يحاولون فيه بالتجربة والخطأ.

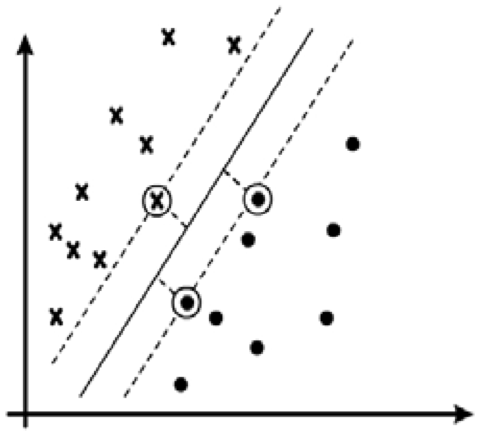
٤ - آليَّات المتجهات الداعمة (Support Vector Machines -SVM)

إنه نوع جديد نسبيا من المصنفات (classifiers)، أثبتت نتائجه تفوقه على كثير من المصنفات الأخرى. وقد قدمه العالم (فابنيك) عام ١٩٩٥. ولكي نفهم المبادئ التي تقوم عليها آليَّات المتَّجهات الدَّاعمة علينا أن ننظر إلى حالة وجود صنفين فقط من الأصناف المراد التفريق بينها. انظر الشكل (٤-٦).



الشَّكل ٤-٦: بيانات صنفين، لكل منهما رمز مختلف.

إن أي خط بين الصنفين سيكون كافيا للفصل بين الصنفين كما في الشكل (٤-٦)، ولكن هناك فاصل سيكون هو الأفضل على الإطلاق، حيث يكون في نصف المسافة بينهما تماما، كما أنَّ له اتجاهًا يكون الفاصل فيه بين الصنفين أكبر ما يمكن، بحيث لو رسمنا خطين متوازيين من ناحيتي «الفاصل الأفضل» سيمسُّ نقاطًا تتبع الصنف الأول ونقاطًا تتبع الصنف الآخَر. انظر الشكل (٤-٧). ولهذه النقاط أهمية كبيرة، إذ هي التي تساهم أساسًا في معادلة «الفاصل الأفضل» ولذلك تسمى «المتجهات الداعمة» (Support Vectors- SV).



الشَّكل ٤-٧: أفضل فاصل بين الصنفين والنقاط التي تمس الحدود (المتجهات الداعمة).

معادلة الحل: لنفرض أن:

X: يمثل متجه الصفات والخصائص التي يمكن قياسها، والمطلوب استخدامها لمعرفة النقطة المقاسَة خصائصها. هل تتبع الصنف الأول أو الصنف الآخَر؟.

y_i: تساوي +١ إذا كانت النقطة المقاسَة تتبع الصنف الأول.

وتساوي -١ إذا كانت النقطة المقاسَة تتبع الصنف الثاني.

n: عدد افراد العينة أي أن $i = 1, \ldots, n$

w: متجه من الثوابت المطلوب الوصول إلى قيمها لمعرفة معادلة الخط الفاصل الأفضل.

b: كمية ثابتة مطلوبة لمعرفة المعادلة الخاصة بالفاصل الأفضل، حيث معادلة الخط (والذي يمكن أن يكون مستوى ذا بُعدٍ أو متعدد الأبعاد) للفاصل الأفضل:

$$w^{T}X - b = 0$$

ويكون الحل كالآتي:

$$w = \sum_{i=1}^{n} a_i y_i \mathrm{X}_i \quad b = y_k - \sum_{i=1}^{n} a_i y_i X_i^{T} \mathrm{X}_k \quad for\ any\ a_k > 0$$

حيث a_i قيم لازمة للحل. وتأخذ القيمة «٠» للمتجه البعيد عن الحدود وله قيمة أكبر من «٠» إذا كان من المتجهات الداعمة. ونقسم مجموعة المتجهات الداعمة s.

ويمكن أن نكوّن دالة التميّز $f(x)$ والتي تكون قيمتها كافية لحسم النقطة إلى أيّ الصنفين تنتمي:

$$f(x) = \sum_{x_i \in s} a_i y_i {X_i}^T X_k + b$$

ومصطلح $X_i \in s$ أى مجموعة المتجهات X_i المنتمية إلى المجموعة s أى مجموعة المتجهات الداعمة.

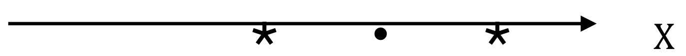
ويتميز هذا الحل بأنه يحمل طابع أنه أفضل فاصل بين الصنفين المراد فصلها. إن الوصول إليه ليس بالتجربة والخطأ، وإنَّما يمكن حسمه بالمعادلات الرياضية.

وهذا الحل يمكن حمله للحالات التي تتداخل فيها نقاط الصنفين، وليس فقط للحالة المثالية التي تناولناها سابقا. ليس فحسب؛ بل يمكن استعماله بطريقة ذكية في حالة الأوضاع الَّتي يستحيل فيها الحل في الفضاء الخطي Linear Space.

خذ مثالا للتوضيح. انظر الشكل (٤-٨)؛ كما ترى في (a) من الشكل (٤-٨)، لا يمكن إيجاد معادلة خطية للفصل بين الصنفين؛ ولكن عند تربيع X (استخدام X^2 مكان X)، أمكن بالرسم إيجاد معادلة خطية للفصل بين الصنفين.

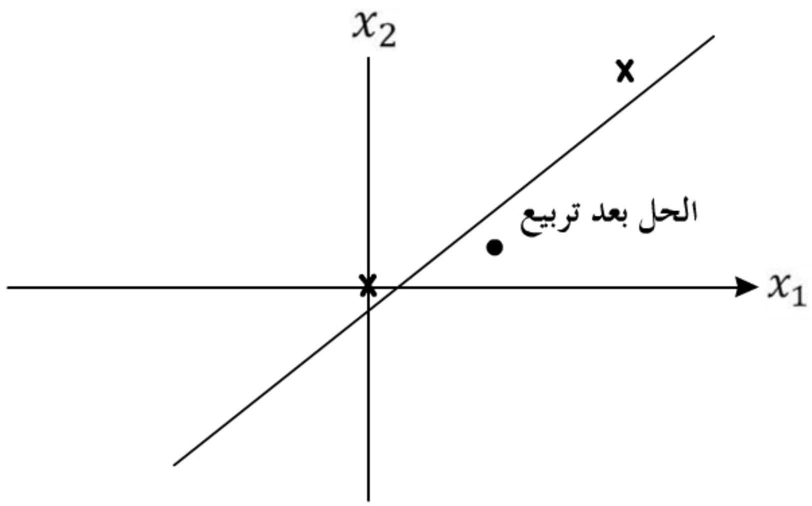
ولتعميم الفكرة فإنَّ الحل يظل الحل الأفضل لو استبدلنا X بـ$\Phi(X)$ حيث $\Phi(X)$ تحويل غير خطي للمتجه X.

هذه العملية قد يكتنفها تعقيدات غير سهلة، كما أنَّ الحَلَّ الأفضل ليسَ مضمونًا في بعض الأحوال، إذ قد نحتاج لتجربة هذه العملية، والتي تسمى نقل المشكلة إلى فضاء آخر أكثر من مرة، مع أنواع مختلفة من هذه الفضاءات، أي الأنواع المختلفة من $\Phi(X)$.



(a)

يتداخل الصنفان (١، ٢) ولا يمكن الفصل بينهما في الفضاء الخطي (a)



(b)

إمكانية إيجاد حلّ إذا ربّعنا القيمة المقاسة x والمستخدمة للفصل بين الصنفين

الشَّكل ٤-٨: كيفية حل مشكلة يستحيل حلها في الفضاء الخطِّيّ.

ومن الملاحظ أنه في الحل $f(x)$ المذكور عاليه لا تظهر x وحدها، ولكن دائما تظهر كالآتي $X_j^T X_k$؛ لذلك سيظهر في الحل بعد عملية الانتقال $\Phi^T(X_j) * \Phi(X_k)$؛ فهل نحن في حاجة لحساب $\Phi(x)$ أصلا؛ الحقيقة لا، وهذا أفضل كثيرا لأن الكمية

$$K(X_j, X_k) = \Phi^T(X_j) * \Phi(X_k)$$

في كثير من الأحيان يكون حسابها أسهل بكثير من حساب $\Phi(X)$ ؛ ولكن لذلك شروط رياضية. وتسمى الدوال التي تخضع للشروط الرياضية هذه (والتي تجعل حسابها ميسورا) بالدوال النواة أو الدوال الجوهرية (Kernel Functions).

٥- نماذج ماركوف المُخَبَّأة (Hidden Markov Models - HMMs)

تُمَثِّل نماذج ماركوف الـمُخَبَّأة مجموعة من النماذج الرياضية الَّتي تُستَخدَم في العديد من التَّطبيقات؛ ومن هذه التَّطبيقات تقنياتُ اللغات الطبيعية. وتُستخدَم هذه النَّماذجُ أساسًا للتعامل مع الظواهر التي تُعرَّف فيها النماذج المراد التعرُّف عليها على أنها سلسلة

من الوحدات المتتابعة. خذ مثالاً لذلك؛ كلمة مثل: «كتب» (مكتوبة ومنطوقة)؛ فإنها مثل أي كلمة تُعرَّف على أنها تتابع من وحدات (كتابية أو صوتية).

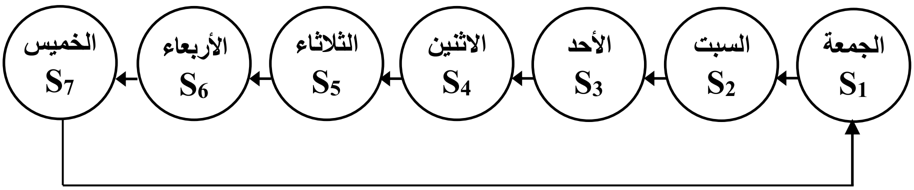
ونبدأ بالتدرج لرسم الحالات (State Diagram)، وتشمل حالات اليقين للاحتمالات المتعددة والمتداخلة.

رسم الحالات (State Diagram)

يتقلب الإنسان عمليا بين حالات كثيرة؛ وفيما يلي بعض الأمثلة التَّوضِيحيَّة.

٥,١ - مثال١:

يحيى الإنسان في الأسبوع بين أيام الجمعة، فالسبت، فالأحد، فالإثنين، فالثلاثاء، فالأربعاء، فالخميس، ثم يعود للجمعة مرة أخرى. وفي كل يوم من هذه الأيام تكون للإنسان حالة مختلفة؛ فإما أن يكون في عمل أو إجازة؛ ويمكن توضيح ذلك في الشَّكل (٤-٩).



الشَّكل ٤-٩: بيان حالات الإنسان لأيام الأُسبوع.

باعتبار أن الإنسان الذي يحيى في يوم الإثنين تُمَثِّله الحالة ٤ أو ٤ State) S_4).

وينتقل الإنسان من حالة إلى أخرى يوميًّا الساعة ١٢ صباحا. وليس في هذا المثال احتمالات وإنما هو مثال للحالات التي يحصل فيها انتقال محدد وغير احتمالي لأنه عند أي مكان محدد على الأرض سيكون الإنسان في حالة محددة من أيام الأسبوع.

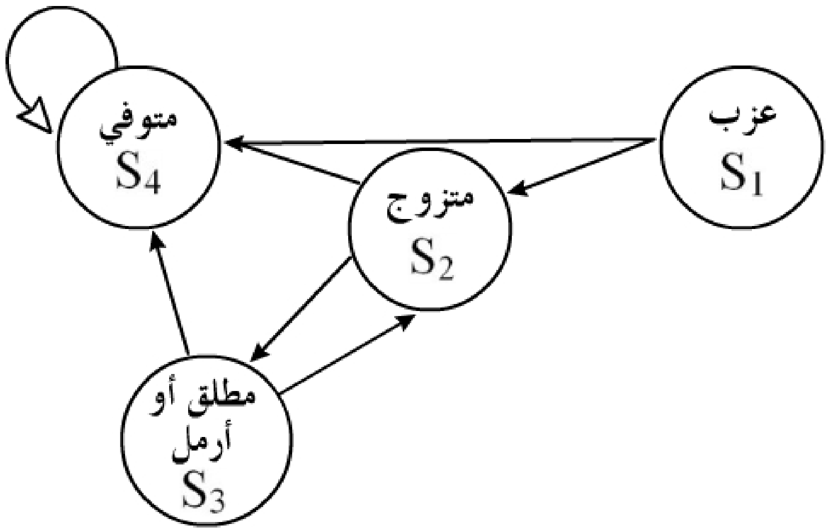
٥,٢ - مثال٢:

حالة الإنسان الاجتماعية؛ فالإنسان يتقلب بين هذه الحالات:

- عَزَب.
- قد يتزوج.

- وقد يطلق.
- متوفَّى.

ويمكن رسم هذه الحالات كما بالشَّكل (٤-١٠)



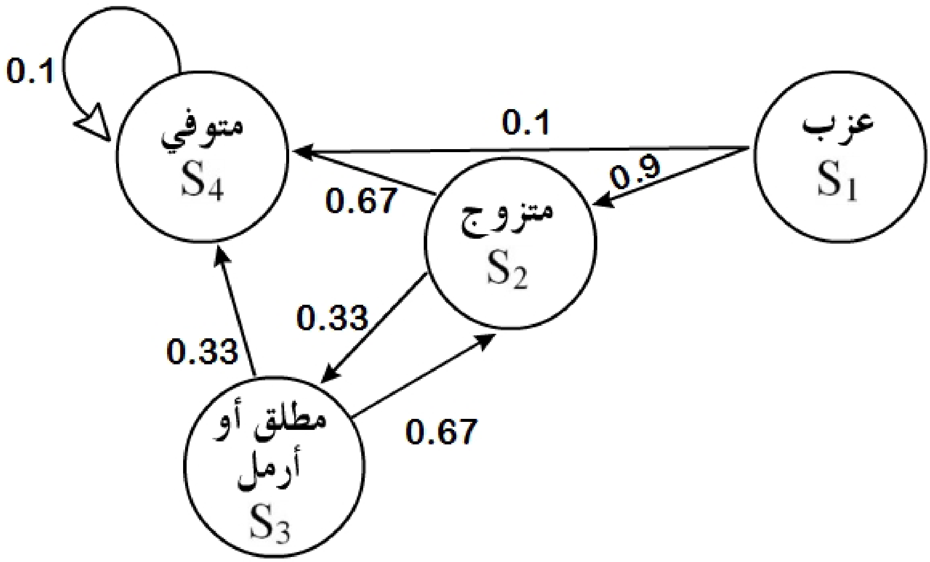
الشَّكل ٤-١٠: الحالات الاحتماليَّة الأربعة S_1, S_2, S_3, S_4.

كما هو مبين في الرسم فإن العَزَبَ إما أن يتزوج وإما أن يظل عزَبًا إلى الوفاة، وكذلك المتزوج إما أن يظل كذلك حتى الوفاة أو ربما يطلق ثم يتوفى أو ربما تتوفى زوجته فيصبح أرملاً، وربما يتزوج أو يبقى كذلك حتى الوفاة.

ولكن في حالتنا هذه ليست الحالات محددة وإنما احتمالية. وبدراسة حياة ١٠٠ حالة في بلد ما وجدنا هذه الأرقام:

- عدد من عاش عمره كله عزَبًا ١٠ أفراد.
- عدد من تزوج ٩٠، ومن طلق أو فقد زوجه ٣٠، أو ثلث من تزوج أى ٣٣٪.
- ٢٠ ممَّن طلق أو ترمل تزوج مرة أخرى، أى الثلثان بنسبة ٦٧٪.

ويمكننا إعادة رسم الشكل (٤-١٠) ليُصبِحَ على النَّحو المبَيَّن في الشكل (٤-١١).



الشَّكل ٤-١١: إحصاء الحالات الاحتماليَّة الأربعة S_1, S_2, S_3, S_4.

في الشكل (٤-١١) يمكن ملاحظة الآتي:

- أن كل حالة يخرج منها سـهم أو أكثر يكون مجموع الاحتمالات للأسهم الخارجية ١.
- في حالة الوفاة يبقى المتوفَّى بالطبع على حاله مهما طال الزمن ولا تتغير حالته؛ ويعبر عن هذا الوضع بالسهم الخارج والداخل لحالة الوفاة، وعليه الاحتمال ١ أي المؤكد.

ويُمكن من خلال المصفوفة A وضع المسألة التي بين أيدينا في شكل رياضي على النَّحو التَّالي:

$$A = \begin{array}{c} \\ \end{array} \begin{matrix} S_1 & S_2 & S_3 & S_4 \end{matrix}$$

$$A = \begin{pmatrix} 0 & 0.9 & 0 & 0.1 \\ 0 & 0 & 0.33 & 0.67 \\ 0 & 0.67 & 0 & 0.33 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{matrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{matrix}$$

وتُلَخَّص المصفوفة A عن المسألة السابقة، حيث نُحَدِّد احتمال الانتقال بين كل حالتين من (S_1, S_2, S_3, S_4) بالرقم المذكور بينهما. مثال؛ الاحتمال أن ننتقل من الحالة ٢ إلى الحالة ٣ يساوى ٠,٣٣ وهكذا.

بعد أن تعرفنا على رسم الحالات، سواء أكان مؤكدًا (Deterministic) أم احتماليًّا (Probabilistic)، سنُحاول أن نعرِضَ مثالاً أكثر تعقيدًا.

٥,٣- مثال٣:

تعال نتصور أن لدينا ٣ أوعية وفي كل وعاء عدة ألوان، ولتكن أربعة ألوان (أحمر، أخضر، أزرق، أصفر)؛ وسنرمز للألوان الأربعة بالرموز (R, G, B, Y). نريد أن نصف عملية معقدة لإخراج الألوان كالآتي:

- سنلقي زهرًا لنحدد بأيّ الأوعية نبدأ (يمكن أن نحول الزهر السداسي إلى ثلاثي إذا اعتبرنا أن رقمي (١، ٢ =>١؛ ٣، ٤ => ٢؛ ٥، ٦ => ٣)؛ وبذلك سنحدد بأي الأوعية نبدأ. ويمكن التعبير الرياضي عن ذلك كالآتي:

$$\pi = (\pi_1, \pi_2, \pi_3)$$

حيث π بمكوناتها الثلاثة تمثل احتمالات البدء لكل وعاء، والشرط أن يكون:

$$(\pi_1 + \pi_2 + \pi_3) = 1$$

- في كل فترة زمنية محددة – ولتكن كل دقيقة- سنلقى الزهر مرة أخرى لنحدد رقم الوعاء القادم، فربما كان نفس الوعاء أو وعاءً آخر. ونعبر عن ذلك بمصفوفة A كما يلي:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

وَتُعّرف A بأنها مصفوفة احتمالات الانتقال بين الأواني؛ وإذا عبَّرنا عن الإناء الذي عليه الدور بالحالة (State) فسوفَ يكون بإمكاننا تسمية A «مصفوفة الاحتمالات الانتقالية» بين الحالات المختلفة المتاحة (Transition Matrix State). مع شرط:

$$(\alpha_1 + \alpha_2 + \alpha_3) = 1$$

وكذلك في بقيَّة الصُّفوف.

- في كل مرة نقف على إناء سوف نمد أيدينا ونأخذ لونًا من ألوانه الأربعة المتاحة عشوائيًّا (بافتراض وجود عدد كبير من كل لون لا يؤثر على النسب بينها أو أنها تعوض ما أُخذ منها حفاظًا على النسب بينها).

ولأن كل إناء يحتوي على نسب مختلفة فإن احتمال خروج أي لون يختلف من إناء لآخر. ويُعبَّر عن ذلك بالمصفوفة B ، حيث:

$$\begin{array}{ccccc} & State_1 & State_2 & State_3 & \\ B = & \begin{pmatrix} b_{11} & b_{21} & b_{31} \\ b_{12} & b_{22} & b_{32} \\ b_{13} & b_{23} & b_{33} \\ b_{14} & b_{24} & b_{34} \end{pmatrix} & & & \begin{matrix} R \\ G \\ B \\ Y \end{matrix} \end{array}$$

وهكذا فإنَّ b_{23} تعني احتمال اللون الأزرق في الحالة الثانية. مع شرط:

$$(\alpha_{11} + \alpha_{12} + \alpha_{13} + \alpha_{14}) = 1$$

وكذلك في بقيَّة الأعمدة.

٥ , ٤ - مثال٤:

بافتراض البدء حتمًا من الإناء الأول، حيث يُسمح فقط للانتقال للإناء اللاحق مع تجهيز الزهر لذلك؛ بمعنى أنه يأخذ القيمة ١ أو ٢ فقط (في هذا المثال يمكن استخدام العملة «ملك = ١،كتابة = ٢»).

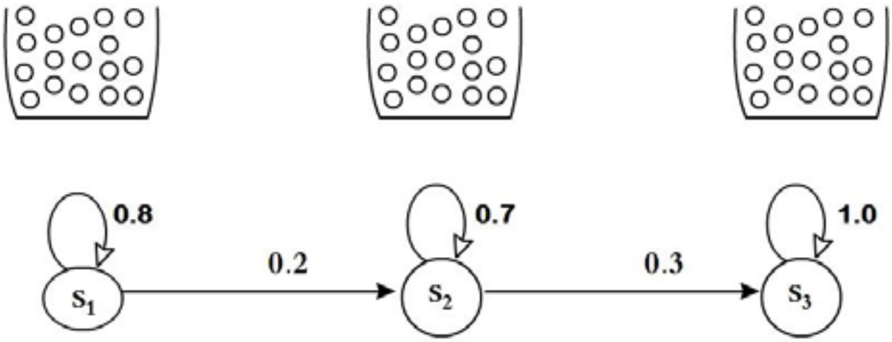
إذا كان ١ ⟵ نفس الحالة (الإناء).

إذا كان ٢ ⟵ الانتقال للإناء الآخر؛ حتى إذا وصل إلى الإناء الأخير توقف الانتقال.

هب أننا بعد وضع الضوابط للمسألة كما أسلفنا وبعد طرح الزهر مرات كثيرة عديدة سجلنا الاحتمالات الآتية:

$$\pi = [١,٠,٠] \;;\; A = \begin{pmatrix} ٠.٨ & ٠.٢ & ٠ \\ ٠ & ٠.٧ & ٠.٣ \\ ٠ & ٠ & ١ \end{pmatrix} \;;\; B = \begin{pmatrix} ٠.٨ & ٠.١ & ٠.١ \\ ٠.١ & ٠.٧ & ٠.١ \\ ٠.٠٥ & ٠.١ & ٠.٣ \\ ٠.٠٥ & ٠.١ & ٠.٥ \end{pmatrix}$$

ويمكن التعبير عن هذه المسألة بالرسم على النَّحو التَّالي:

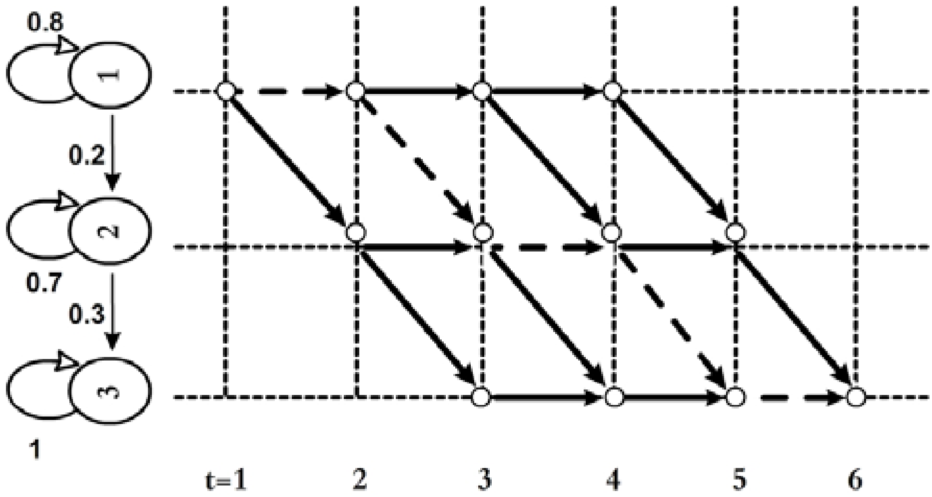


$$\begin{matrix} R \\ G \\ B \\ Y \end{matrix} \begin{pmatrix} ٠.٨ & ٠.١ & ٠.١ \\ ٠.١ & ٠.٧ & ٠.١ \\ ٠.٠٥ & ٠.١ & ٠.٣ \\ ٠.٠٥ & ٠.١ & ٠.٥ \end{pmatrix}$$

<u>وإليك المسألة</u>؛ هب أننا حصلنا على ترتيب الألوان من اليسار إلى اليمين، ولا نعرف أيَّ الألوان خرجَ من أيّ وعاء؛ كل ما نعرفه أن اللون الأول (R في حالتنا) أُخذ حتمًا من الإناء الأول S_1 وأن اللون الأخير (Y في حالتنا) أُخذ من الإناء الأخير S_3.

Time:	1	2	3	4	5	6	الزمن:
Colors:	R	R	G	G	B	Y	الألوان:

نريد أن نعرف ترتيب الأواني، إذ في معرفتها حلٌّ للمشكلة. لو فكرتَ قليلا لعلمتَ أن هناك احتمالات كثيرة جدًّا لترتيب تلك الأواني، وهي تؤدي إلى نفس ترتيب الألوان المذكورة عاليه. ويمكن رسم المسارات الممكنة كما في الشكل (٤-١٢).



الشَّكل ٤-١٢: عرض لمسارات الحل الممكنة باستعراض الحالات المتاحة مع الزمن.

لو تتبعت الشبكة المرسومة في الشكل (٤-١٢) لأمكنك تتبُّع عشرة مسارات مختلفة، ولو اخترنا أحد هذه المسارات (المؤشر في الشكل)، كيف نحسب احتمالية هذا المسار كمثال؟

احتمالية المسار المختار للدراسة = P(of the selected path)

$$= \underbrace{\pi_1 * P(R/s_1)}_{t=1 \text{ الزمن}} * \underbrace{P(s_1/s_1) * P(R/s_1)}_{t=2} * \underbrace{P(s_2/s_1) * P(G/s_2)}_{t=3} *$$

$$\underbrace{*P(G/s_2) * P(s_3/s_2)}_{t=4} \underbrace{*P(B/s_3) * P(s_3/s_3)}_{t=5} \underbrace{*P(Y/s_3)P(s_2/s_2)}_{t=6}$$

$$= (\pi_1 * b_{11}) * (a_{11} * b_{11}) * (a_{12} * b_{22}) * (a_{22} * b_{22}) * (a_{23} * b_{33}) * (a_{33} * b_{34})$$

= (1 * 0,8) * (0,8 * 0,8) * (0,2 * 0,7) * (0,7 * 0,7) * (0. 3 * 0,3) * (1 * 0,5)

= 0,00158

وهكذا لو حسبنا المسارات العشرة سوف نجد أن أحد هذه المسارات هو الأعلى احتمالا؛ ويُرشَّح هذا المسار لأن يصف ترتيب الأواني التي تعاملنا معها عبر ٦ وقفات زمنية. وحتى يتَّضحَ ما قمنا به فإننا عند كل وقفة زمنية نحسب:

(احتمال أن نصل إلى الحالة (الإناء) التي وصلنا إليها) * (احتمال خروج اللون الذي خرج من الإناء الذي نقف عنده)

ثم نكرر ذلك عبر الوقفات الزمنية كلها.

ويسمى هذا النموذج الرياضي «نموذج ماركوف المـُخَبَّأ» (HMM). ولهذا النموذج الرياضي ٣ مسائل:

المسألة الأولى:

إذا توافرت كميات مناسبة من المشاهدات المتتابعة

$$O = o_1, o_2, \ldots o_T$$

وتوفر كذلك نموذج HMM، ويعرف رياضيًّا كالآتي:

$$\lambda = (\pi, A, B)$$

فما هو احتمال أن تنتمي المشاهدات إلى النموذج الرياضي HMM؟

ويعبر عن ذلك بـ $P(O/\lambda)$

المسألة الثانية:

المعطى: المشاهدات المتابَعة

$$O = o_1, o_2, \ldots o_T$$

وكذلك نموذج HMM الرياضيّ:

$$\lambda = (\pi, A, B)$$

والمطلوب معرفة تتابع الحالات (State Sequence) الأكثر احتمالا؛ وتسمى هذه المشكلة فيتربي Viterbi.

المسألة الثالثة:

المعطى: المشاهدات O. والمطلوب: تقدير قيم معاملات النَّموذج الرِّياضِيّ $\lambda = (\pi,A,B)$.

أي تقدير قيم a_{ij} ، b_{ik} ، π_i، والتي تجعل الكمية $P(O/\lambda)$ أعلى ما يمكن.

وقبل الخوض في الحسابات المرتبطة بنماذج ماركوف الـمُخَبَّأة، تعالوا نراجع بعض التعريفات.

$$O = o_1, o_2, o_3, o_4, \ldots o_T$$

تعني سلسلة المشاهدة O، والتي تتكون من عدد T مشاهدة؛ تعني في حالتنا سلسلة الألوان المتتابعة والتي من المفترض أن نبحث فيها عن سلسلة الأواني التي أخرجنا منها هذه الألوان المتتابعة.

عدد الحالات (الأواني) في النموذج تحت الدراسة $=N$

عدد الألوان التي يمكن استخراجها من أي أناء $=M$

رمز لكل حالة أو أناء $S = S_1, S_2, S_3, \ldots S_N$

مصفوفة الانتقال بين الحالات، حيث a_{ij} يمثل المكون $A= \{b_{ik}\}$

رقم i , j في المصفوفة، وهو يمثل احتمال الانتقال من الحالة (i) إلى الحالة (j)

مصفوفة ربط الألوان أو الرموز (k) المنبعثة من الحالة (i) $B = \{b_{ik}\}$

(أو الإناء). و b_{ik} تعنى احتمال إخراج اللون k من الحالة (الإناء i).

احتمالات البدء $\pi = \pi_1, \pi_2, \ldots \pi_N$

حيث π_j تعني احتمال البدء بالحالة (j).

وفي كثير من التطبيقات نفرض على النموذج البدء بالحالة الأولى، وهو ما يعني أنَّ $\pi_1 = 1$ وبقيَّة حالات البدء = ٠؛ وفي هذه الحالة يكون $\pi = (1,0,0,\ldots 0)$.

حلّ المسألة الأولى: خوارزم «للأمام-للخلف»Forward-Backwar Algorithm

لنبدأ بتعريف

$$\propto_t^{(i)} = P_r(o_1, o_2, ..., o_t, \quad i_t = s_i / \lambda)$$

أي احتمال مرور سلسلة المشاهدات من $o_1,o_2,...o_t$، ومع البدء بالحالة الأولى؛ هذا باعتبار أن لدينا نموذج ماركوف مُخبّأً بعينه λ، حيث i_t تمثل رقم الحالة i عند الزمن t. ويتكون الخوارزم «للأمام ثم الخلف» من ثلاث خطوات:

الخُطوة ١: خطوة البدء

$$\propto_t^{(i)} = \pi_i b_j(o_1), \qquad 1 \le i \le N$$

حيث $\propto_t^{(i)}$ = تحتوي على مجموع احتمالات المسارات من البدء إلى الحالة (i) في زمن (t).

الخُطوة ٢: خطوة التكرار

$$\text{For} \quad t = 1, 2, ... T-1, \qquad 1 \le j \le N$$

$$\propto_{t+1}^{(j)} = [\sum_{i=1}^{N} \propto_{t(j)} . a_{ij}] * b_j(o_{t+1})$$

حيث $b_j(o_{t+1})$ تعني احتمال أن يكون اللون (أو الرمز) عند الزمن t+1 خارجا من الإناء (j).

$\sum_{i=1}^{N}$ تعني مجموع المسارات الواردة بتغيير الحالة رقم (i) من ١ إلى N.

الخُطوة ٣: خطوة الانتهاء

$$P(O/\lambda) = \sum_{i=1}^{N} \propto_T (i)$$

حيثُ $P(O/\lambda)$ تعني احتمال انتهاء سلسلة المشاهدات O إلى النموذج λ، وتحسب كمجموع احتمالات المسارات المحتملة من البدء إلى النهاية عند الزمن T.

وفي كثير من التطبيقات يكون لزامًا علينا أن نبدأ بالحالة الأولى وننتهي بالحالة الأخيرة N؛ وعندئذ:

$$P(O/\ \lambda) = \propto_T^{(N)}$$

ويسمى النموذج في هذه الحالة نموذج **الشمال-يمين** (Left-right model).

حلّ المسألة الثانية: خوارزم فيتربي:

يتكون هذا الخوارزم من أربع خطوات:

الخُطوة ١: خطوة البدء

$$\delta_1(i) = \pi_i * b_i(o_1) \qquad 1 \le i \le N$$

$$\psi_t(i) = 0$$

حيث $\delta_1(i)$ تحتوي على احتمال المسار الأعلى احتمالا من البدء إلى الحالة (i) في زمن (t)، و $\psi_t(i)$ تحتوى على رقم الحالة التي انتقلنا منها إلى الحالة (i) على المسار الأعلى احتمالا في الزمن (t).

الخُطوة ٢: خطوة التكرار

الوقت t يتغير من $2 \le t \le T$، والحالة (j) تتغير من $1 \le j \le N$

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i) * a_{ij}] * b_j(o_t)$$

$$\psi_t(j) = \operatorname{argmax}_{1\le i\le N} [\delta_{t-1}(i) * a_{ij}]$$

حيث $b_j(o_{t+1})$ تعني احتمال أن يكون اللون عند الزمن t خارجا من الإناء (أو الحالة) رقم (i).

و $\max_{1\le i\le N}[]$ تعني أننا نحسب أعلى قيمة لما بين الأقواس [ ] بتغيير قيمة (i).

و $\operatorname{argmax}_{1\le i\le N}[]$ تعني أننا نحتفظ برقم (i) الذي أعطى أعلى قيمة لما بين الأقـواس [ ]، وليس قيمة الحسبة نفسها

الخُطوة ٣: خطوة الانتهاء

$$P^* = \max_i \, [\delta_T(i)]$$

$$i_T^* = \operatorname{argmax}_i \, [\delta_T(i)]$$

حيث *P تعنى احتمال انتماء سلسلة ألوان متتابعة O إلى النموذج $\lambda = (\pi, A, B)$ على أساس حساب المسار الأعلى احتمالا.

و i_T^* هو رقم الحالة على المسار الأعلى احتمالا عند الانتهاء بالزمن T.

الخُطوة ٤: خُطوة معرفة المسار الأكثر احتمالاً

For t = T-1, T-2,…2,1

أي بتراجع الزمن

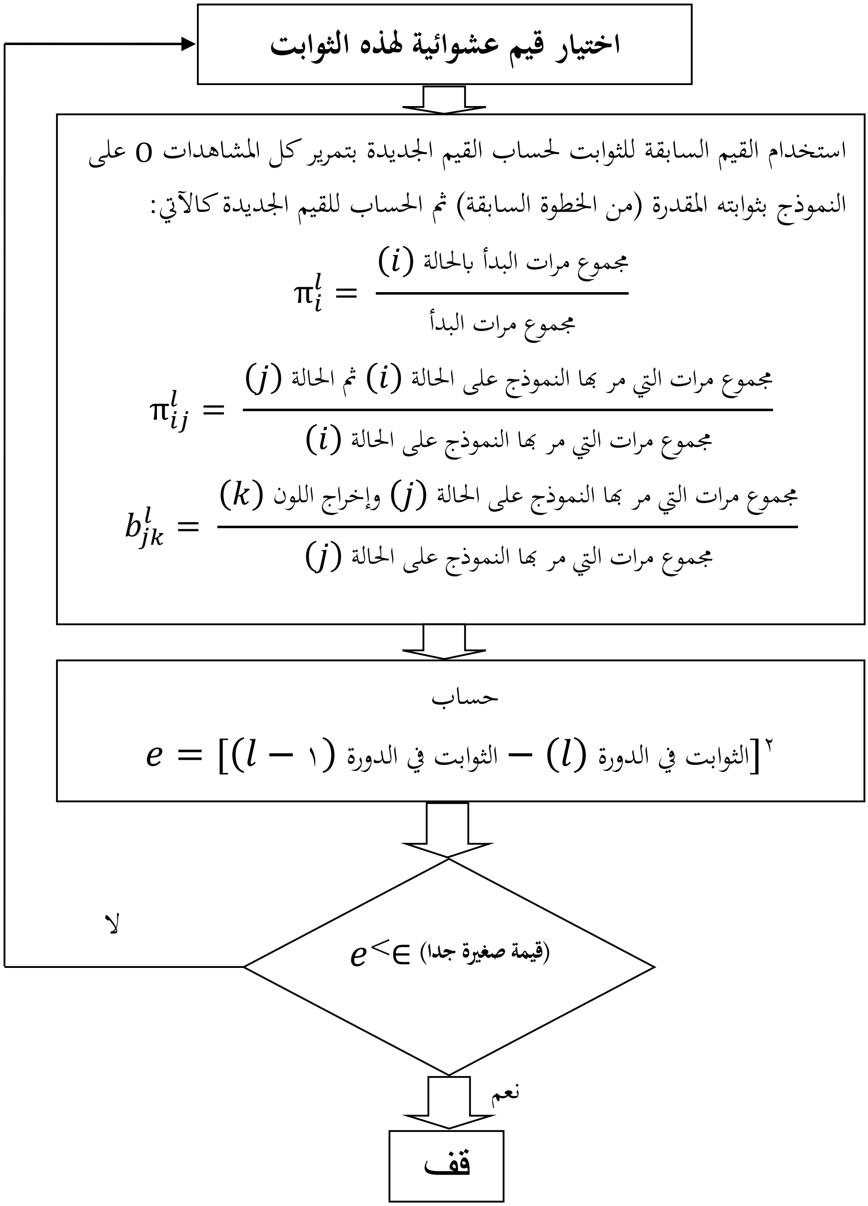
$$i_t^* = \psi_{t+1}(i_{t+1}^*)$$

حساب i_t^* تعنى «بالتراجع» يمكن الحساب عند كل زمن t الحالة (i) التى تقع على المسار الأعلى احتمالا.

حلّ المسألة الثالثة: تقديرات «بوم- ولاش» لثوابت نموذج ماركوف الـمُخَبَّأة $\lambda=(\pi, A, B)$

أي حساب قيم: b_{ik}, a_{ij}, π_i

ويُمكن القيام بذلك على النَّحو التَّالي:


اختيار قيم عشوائية لهذه الثوابت
استخدام القيم السابقة للثوابت لحساب القيم الجديدة بتمرير كل المشاهدات O على النموذج بثوابته المقدرة (من الخطوة السابقة) ثم الحساب للقيم الجديدة كالآتي:
$\pi_i^l =$ مجموع مرات البدأ بالحالة (i) / مجموع مرات البدأ
$\pi_{ij}^l =$ مجموع مرات التي مر بها النموذج على الحالة (i) ثم الحالة (j) / مجموع مرات التي مر بها النموذج على الحالة (i)
$b_{jk}^l =$ مجموع مرات التي مر بها النموذج على الحالة (j) وإخراج اللون (k) / مجموع مرات التي مر بها النموذج على الحالة (j)
حساب
$e = [($الثوابت في الدورة (l) − الثوابت في الدورة $(l-١))]^{٢}$
$e < \in$ (قيمة صغيرة جدا)
لا
نعم
قف

وفي ختام هذا الفصل تَجدُرُ الإشارةُ إلى التطور الهائل في مجال استخدام الشبكات العصبية في الأبحاث الخاصة بمجال حوسبة اللغات الطبيعية. لقد تطورت الأشكال والأنماط لهذه الشبكات العصبية تطورا هائلا وأعطت نتائج في معظم الحالات أفضل بكثير من تلك النتائج التي كنا نحصل عليها بالطرق التقليدية. إلا أنه من الملاحظ أن الطرق التقليدية تتفوق عندما يكون حجم البيانات المخصصة للتدريب قليلًا نسبيا. وحتى في هذه الحالة هناك نماذج ظهرت سبق أن تدربت على بيانات كثيرة متوفرة؛ ولكن لمهام مختلفة أو للغة أخرى. وعندئذ يبدؤون تدريب هذه النماذج سالفة التدريب على القليل من البيانات المتاحة، فإذا بها تعطي نتائج ممتازة. ستكون السنوات القادمة مليئة بإنجازات هائلة في مجال حوسبة اللغات الحية بما يقربها من المستوى البشري المعجز. وستكون هذه من الفتوحات العلمية التي مَنَّ الله علينا بها، وسيتجلى تأثير ذلك في كل مناحى الحياة.

ببليوجرافيا مرجعيَّة

1. Balivada, L. K. & Raju, K. P. (2012): Optimization Techniques of Viterbi Algorithm: Performance Analysis of Different Algorithms. Lambert Academic Publishing.
2. Cristianini, N. & Shawe-Taylor, J. (2000), An Introduction to Support Vector Machines and other kernel-based learning methods, Cam-bridge University Press.
3. Deng, L.; Liu, Y. (2018). Deep Learning in Natural Language Processing. Springer.
4. Fraser, A. M. (2008): Hidden Markov Models and Dynamical Systems. SIAM.
5. Haykin, S. (1999), Neural Networks: A Comprehensive Foundation, Prentice Hall.
6. Kaleli, C. & Polat, H. (2010): NAÏVE BAYESIAN CLASSIFIER-BASED PRIVATE RECOMMENDATIONS: PRIVACY-PRESERVING NAÏVE BAYESIAN CLASSIFIER-BASED COLLABORATIVE FILTERING. LAP Lambert Acad. Publ.
7. Karwowski, W. (2019). Intelligent Human Systems Integration 2019. Springer.
8. Kubat, M. (2012), Machine Learning, illustrated, Eleven Learning.
9. Kulkarni, A.; Shivananda, A. (2019). Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python. Apress.
10. Mamon, R. S. & Elliott, R. J. (2010): Hidden Markov Models in Finance. Springer.
11. Neamat El, G.; Yee, S. (2018). Computational Linguistics, Speech and Image Processing for Arabic Language. World Scientific.

12. Rish, I. (2001). “An empirical study of the naive Bayes classifier”. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. Availabl online here: http://www.research.ibm.com/ people/r/rish/papers/RC22230.pdf
13. Russell, J. & Cohn, R. (2012): Viterbi Algorithm. Book on Demand.
14. Sharp, B.; Sedes, F.; Lubaszewski, W. (2017). Cognitive Approach to Natural Language Processing. Elsevier.
15. Srinivasa-Desikan, V. (2018). Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing.
16. Vojislav, K. (2001), Learning and Soft Computing, Support Vector Ma-chines, Neural Networks and Fuzzy Logic Models, The MIT Press, Cambridge, MA.
17. Zizka, J.; Darena, F.; Svoboda, A. (2019). Text Mining with Machine Learning. Taylor & Francis Group.

الفصل الخامس
نمذَجة اللُّغة

د. مُحسِن رَشوان

١- النَّحو العدديّ.

٢- التَّنعيم.

٣- موضوعات تساعد على تحسين النَّحو العدديّ.

٤- تقويم قوة النَّحو العَدَدِيّ.

٥- مجالات الإفادة من النَّحو العَدَدِيّ.

٦- أفكارٌ بحثيَّة لأطرُوحاتٍ علميَّةٍ مُستَقبليَّة.

تمهيد

اللُّغات الحية في الحقيقة معقَّدة بما فيه الكفاية لتلبية حاجة الإنسان في التَّعبير عن مشاعره وأفكاره المتجدِّدة، إذ يمكن للإنسان أن يعبِّرَ عن معنى يجول في خاطره بعددٍ كبير جدًا من الجُمَل الَّتي تؤدِّي نفسَ المعنى. وربما تختلف عن بعضها في الدِّقة والبلاغة، والمشاعر المحيطة بالمعنى... إلخ. وهذا يجعل وضعَ إطار رياضيٍّ دقيقٍ للتعبير عن فهم المتحدث وقصده أمرًا بالِغَ الصُّعُوبة - إن لم يكُن مُستَحيلاً - في الوقت الحالي؛ وفي الوقت ذاته لا نستطيعُ الاستغناءَ عن نمذجة اللُّغة الَّتي تُوَجِّهُ التِّقنيات اللُّغَويَّة إلى تحقيق أهدافها المنشودة في مجالاتٍ مُتَعَدِّدة، كالتَّعرُّف الآليّ على الكلام المكتوب أو المنطوق.

وعلى سبيل المثال، في مجال التَّعَرُّف على الكلام المنطوق، لو افتَرَضنا أنَّ المتَحدِّث نطقَ جملةً تحتوي على ١٠٠ فُونيما متتاليًا (حوالي ١٠-١٢ كلمة متصلة) - آخذينَ في الاعتبارِ أن أدقَّ الأنظمة التي تتعرَّف على الكلام المنطوق لا يتجاوز متوسط دقَّتها ٨٠٪ لكل فونيم على حدة - فإنَّ درجة دِقَّة التّقنية على مستوى الجمل إذا تخلَّينا عن استخدام النَّموذَج اللُّغَوِيّ ستكونُ على النَّحو المبَيَّن في الجدول التَّالي:

عدد الفونيمات المكونة للكلمة أو الجملة بافتراض متوسط دقة ٨٠٪ لكل فونيم	
عدد الفونيمات	**دقَّة التَّعَرُّف**
١	٪٨٠
٢	٪٦٤ = ٢^(٪٨٠)
٣	٪٥١,٢ =٣^(٪٨٠)
١٠ (متوسط الكلمة)	٪١٠,٧ = ١٠^(٪٨٠)
٥٠ (جملة قصيرة)	٪٠,٠٠١٤ = ٥٠^(٪٨٠)
١٠٠ (جملة متوسطة)	٪٠ ≈ ١٠٠^(٪٨٠)

الجدول ٥-١: دقَّة الكلمات والجمل بدون نموذج لغويّ.

ووفقًا لهذه النَّتائج، سيُؤَدِّي الاستغناء عن النَّموذَج اللُّغَوِيّ إلى نتائج ليست ذات قيمة، وبالتَّالي ستُصبِحُ تقنية التَّعَرُّف على الكلام المنطوق عديمة الفائدة بخُلُوِّها من هذا النَّموذَج. أمَّا إذا اعتمدنا عليه فإنَّ الكلمات العربية في صُورتها المفردة ستتحرَّك

من ٧,١٠٪ إلى أكثر من ٩٠٪ (في ظروف تسجيل مناسبة)، بمساعدة النَّماذِج اللُّغَوِيَّة باعتبارِها مجموعة من المعلومات الرِّياضِيَّة الموضوعة في قالبٍ رياضِيّ؛ وبعبارةٍ أخرى، تُساعد «نمذجة اللغة» (Language Modeling) في تحقيق الفائدة من تقنيات اللُّغات.

ونستطيعُ التَّمثيل على ذلك بتحليل المقطع الصَّوتيّ «ذَهَبَ إلى»، حيث تحتملُ اللَّفظة «إلى» أن تكونَ ١. «آلي»، أو ٢. «آلا»، أو ٣. «إلى». ونستطيعُ أن نستَدِلَّ على الاحتمال الأقرب إلى الصَّواب بتحليل تتابُع هذه الكلمات في سياقاتها اللُّغَوِيَّة؛ وبافتراض أنَّنا قُمنا بتحليل كلمة «ذَهَبَ» وتعرَّفنا عليها بشكلٍ صحيح، فإنَّنا سنجد أنَّ كلمة «إلى» هي الأكثرُ التصاقًا بها، مايعني أنَّ الاحتمالَ الثَّالثَ أقربُ إلى الصَّواب.

١- النحو العدديّ (N-gram)

هناك بعضُ الطُّرُق الَّتي تُستَخدَم في توجيه تقنيات اللُّغة وتطبيقاتها إلى الاحتمال الأقرب إلى الصَّواب من النَّاحية اللُّغَوِيَّة؛ ويُعَدُّ النَّحو العددِيّ «N–gram» أوسعَ هذه الطُّرُق انتشارًا وأكثرها استخدامًا. وسنُحاول الوُقُوفَ -فيما يلي- على أهمِّيَّة النَّحو العدديّ ودوره في تقنيات اللُّغة، مُقَدِّمِينَ له بالحديث عن الاحتمالات [٥،٤، ١٤].

١,١- حساب الاحتمالات والاحتمالات الشَّرطيَّة

إذا كانت لدينا مُدوَّنةٌ لُغَوِيَّةٌ تضُمُّ مليون كلمة، وكانت إحدى كلماتها قد ورَدَت ١٠٠٠ مرة، فإنَّنا نستطيعُ تحديدَ احتمال ورُود هذه الكلمة في وثيقة تتشابهُ مادَّتُها مع مادَّة المدونة اللُّغَوِيَّة باستخدام المعادَلة التَّالية:

$$\text{احتمال الورود} = \frac{\text{عدد مرات ورود الكلمة في المدونة}}{\text{عدد كلمات المدونة كلها}} = P(w)$$

$$P(w) = \frac{١٠٠٠}{١٠٠٠,٠٠٠} = ٠,٠٠١ = ٪٠,١$$

أي واحد في الألف.

حيث ترمُز w إلى الكلمة، (اختصارًا لـ Word)،

وترمز $P(w)$ إلى احتمال الوُرُود، (اختصارًا لـ Probability of Word).

دعنا نعرف الاحتمالات المشروطة في هذا المثال: ورد في القرآن الكريم كله عدد ٧٧٩٣٤ كلمة، وكانت كلمة «الله» هي الأكثر ورودًا فيه، وجاءت هذه اللفظة الكريمة ٢٧٠٧ مرة، فكانت مرفوعة في ٩٨٠ مرة، ومنصوبة في ٥٩٢ مرة، ومجرورة في ١١٣٥ مرة. فلو سألنا عن احتمال ورود كلمة الله في القرآن الكريم كله ستكون الإجابة:

$$P(\text{الله})=\frac{٢٧٠٧}{٧٧٩٣٤}=٪٣,٤٧$$

بينما لو سألنا عن كلمة القرآن مرفوعة في القرآن كله تكون الإجابة:

$$P(\text{اللهُ})=\frac{٩٨٠}{٧٧٩٣٤}=٪١,٢٦$$

ماذا لو سألنا هذا السؤال: ما احتمال ورود كلمة (الله) مرفوعة منسوبة إلى كل كلمات (الله) في القرآن الكريم؟ أو بعبارة أخرى: ما احتمال ورود كلمة (الله) مرفوعة بشرط نسبها إلى كلمة (الله) في القرآن كله؟ سيكون التَّعبير رياضيًّا على هذا النَّحو:

$$P(\text{اللهُ} / \text{الله})$$

وتعني الشَّرطة المائلة في التَّعبير الرِّياضيّ السَّابق أنَّ شرطَ حساب احتمال ورود كلمة (الله) مرفوعة هو ورود كلمة (الله) أيًّا كان تشكيلها. ويكون حسابها كالآتي:

$$P(\text{اللهُ} / \text{الله})=\frac{٩٨٠}{٢٧٠٧}=٪٣٦,٢$$

١,٢ - النَّحو العدديّ الأحاديّ (Uni-gram)

بعد أن قدَّمنا فكرةَ الاحتمالات الشَّرطيَّة، يُمكنُنا أن نُقَدِّمَ فكرةً عن النَّحو العدديّ. كما أسلفنا في المقدمة أنَّنا في حاجة ماسَّة إلى معلومات عن اللُّغة وعن تردُّد كلماتها وترابطها معًا، لنتمكَّنَ من دعم الحل الصحيح في تقنيات كثيرة من تقنيات اللُّغات الحية. والواقعُ أنَّ ظُهُور النَّحو العدديّ الاحتماليّ في الخمسينيَّات من القرن العشرِين باعتبارِه مسارًا إحصائِيًّا يُستَخدَم في مُعالَجَة اللُّغات الحيَّة قد لاقى عزوفًا من قبل اللُّغَويِّين في ذلك الوقت نتيجةَ ما أفرزَته نظريَّات اللُّغَوِيّ الأمريكيّ نَعُوم تشومسكي من نقدٍ لهذا المسار. ولكن بعد أن نجحت شركة IBM في السبعينيَّات من العودة إلى

النَّحو العدديّ بنجاح، اتَّجَهَ الباحثون في تقنيات اللُّغات الحية إلى الاستعانة به، حتى غدا أساسا لا غنى عنه لمطوري هذه التقنيات.

دعنا نأخذ مثالاً مطولاً لفهم النَّحو العدديّ [أو الإحصائِيّ] N-gram. لو تصوَّرنا أنَّ لدينا مدونةً لغوية مُبَسَّطة تتكون من هاتَين الجملتَين:

«ذهب محمد إلى المدرسة»
«حين وصل محمد إلى المدرسة قابل زميله أحمد»

عدد الكلمات في هذه المدونة المبسطة ١٢ كلمة؛ وبإضافة رمز لبداية جملة (مَرَّتين)، ورمز نهاية جملة (مَرَّتين)، وكأنهما كلمتان مضافتان لمفردات المدونة، يكون عدد الكلمات ١٦ كلمة. أي: عدد مفردات المدونة ١٦ مفردة (١٢ + بدايتين لجملتين ونهايتين).

وقبل الشُّروع في توضيح مفهوم النَّحو العددِيّ، نودُّ أن نُشِيرَ إلى قيام عالِمِ الرِّياضيَّات الرُّوسيّ أندريه ماركوف (١٨٥٦-١٩٢٢) بوضع نموذجٍ رياضيٍّ مبسطٍ للتَّنبُّؤ بالمستقبل بالاستعانة فقط ببضع خطوات من الماضي. وسوف نستفيد من تبسيطه الرياضيّ فيما يلي:

لنحسب للمدونة المَبَسَّطة السَّابقة (والَّتي لا يتعدَّى محتواها ١٦ كلمة) حسابات تدخل في مفهوم النَّحو العدديّ:

أولاً: تُعرَفُ الدَّرجة الأولى في النحو العدديّ بـ «النحو الأحادي uni-gram» أو 1-gram؛ وفيه نحسب فقط احتمالية تكرار كل كلمة بصرف النظر عن ما قبلها أو ما بعدها، على النَّحو المبَيَّن في الجدول التَّالي:

م	مُفردات المدونة	التَّرَدُّد (الوُرُود)	النَّحو الأحاديّ	النِّسبة
١	بداية جملة	٢	٢/١٦ = ١/٨	٠,١٢٥
٢	ذهب	١	١/١٦	٠,٠٦٢٥
٣	محمد	٢	٢/١٦ = ١/٨	٠,١٢٥
٤	إلى	٢	٢/١٦ = ١/٨	٠,١٢٥
٥	المدرسة	٢	٢/١٦ = ١/٨	٠,١٢٥

م	مُفردات المدونة	التَّرَدُّد (الوُرُود)	النَّحو الأحاديّ	النِّسبة
٦	حين	١	١٦/١	٠,٠٦٢٥
٧	وصل	١	١٦/١	٠,٠٦٢٥
٨	قابل	١	١٦/١	٠,٠٦٢٥
٩	زميله	١	١٦/١	٠,٠٦٢٥
١٠	أحمد	١	١٦/١	٠,٠٦٢٥
١١	نهاية الجملة	٢	٨/١ = ١٦/٢	٠,١٢٥
	المجموع	١٦	١,٠٠	١,٠٠

الجدول ٥-٢: حسابات النَّحو الأحاديّ لمفردات المدَوَّنة المبَسَّطة.

١,٣ – النَّحو العدديّ الثنائيّ (Bi-gram)

يمكن الارتقاء درجةً وحساب النَّحو العدديّ إذا نظرنا خلفنا لكلمة واحدة، واستعنَّا بهذه المعلومة لحسابات المستقبل. فبالنَّظر إلى الجدول رقم (٥-٢) سنُلاحظُ أنَّنا نضَع في حساباتنا (بداية الجملة) و (نهاية الجملة). ويسمَّى هذا بالنَّحو الثُّنائيّ [٥، ٢٤].

		الكلمة السابقة										
		بداية جملة	ذهب	محمد	إلى	المدرسة	حين	وصل	قابل	زميله	أحمد	نهاية جملة
١	بداية جملة											
٢	ذهب	١										
٣	محمد		١					١				
٤	إلى			٢								
٥	المدرسة				٢							
٦	حين	١										
٧	وصل						١					
٨	قابل					١						

		الكلمة السابقة										
		بداية جملة	ذهب	محمد	إلى	المدرسة	حين	وصل	قابل	زميله	أحمد	نهاية جملة
٩	زميله								١			
١٠	أحمد									١		
١١	نهاية جملة					١					١	
	المجموع	٢	١	٢	٢	٢	١	١	١	١	١	

الجدول ٥-٣: النحو الثنائي للمدَوَّنة.

الاحتمال الشرطي للنحو الثنائي $P(w_n/w_{n-1})$	عدد ورود الكلمة $C(w_{n-1})$	عدد ورود الكلمتين معاً $C(w_n,w_{n-1})$	$\frac{C(w_n,w_{n-1})}{C(w_{n-1})}$	$P^*(w_n/w_{n-1}) = \frac{C(w_n,w_{n-1})+0.01}{C(w_{n-1})+121*0.01}$
P(ذهب/ بداية جُملة)	٢	١	٠,٥	٠,٣١٥
P(محمد/ ذهب)	١	١	١	٠,٤٥٧
P(إلى/ محمد)	٢	٢	١	٠,٦٢٦
P(المدرسة/ إلى)	٢	٢	١	٠,٦٢٦
P(نهاية جُملة/ المدرسة)	٢	١	٠,٥	٠,٣١٥
P(حين/ بداية)	٢	١	٠,٥	٠,٣١٥
P(وصل/ حين)	١	١	١	٠,٤٥٧
P(قابل/ المدرسة)	٢	١	٠,٥	٠,٣١٥
P(زميله/ قابل)	١	١	١	٠,٤٥٧
P(أحمد/ زميله)	١	١	١	٠,٤٥٧
P(نهاية جُملة/ أحمد)	١	١	١	٠,٤٥٧
اى تتابع ثنائى لم يرد عاليه	لو كانت: $C(w_{n-1}) = 1$	٠	٠	٠,٠٠٤٥
	لو كانت: $C(w_{n-1}) = 2$	٠	٠	٠,٠٠٣١

الجدول ٥-٤ : النحو الثنائي للمدَوَّنة. العمود الثالث محسوب فيه النحو الثنائي بدون مراعاة

لـ OOV، والعمود الأخير محسوب فيه النحو الثلاثي بعد مراعاة OOV.

وهكذا، لو أردنا درجةً أخرى من نحوٍ أعمق فبإمكاننا أن نلجأ للنَّحو الثُّلاثيّ (3-gram)، وعندئذ يكون مثلاً:

$$P\,(\text{المدرسة / محمد، إلى}) = \frac{٢}{٢} = ١$$

وعليه، يمكنُ حسابُ النَّحو الرُّباعيّ والخُماسيّ... إلخ.

والآن، نريد أن نقف عند مشكلة خطيرة في هذا الطرح، ألا وهي: ماذا نفعل مع الكلمات الَّتي لم ترد في سياق المدَوَّنة؟ سيكون احتمال وُرُودِها صفرًا، وهذا يتغير كثيرا إذا حسبنا أن ما لم نره في المدونة يكون احتمال وُرُودِه صفرا [٣، ٢١].

مثال: إذا قابلتنا عبارة (ذهب أحمد إلى المدرسة)، وأردنا الاستفادة من المدونة السابقة في استنباط نتائج مفيدة:

$$P(\text{ذهب .أحمد}) = ٠$$

سوف نجد أنها تساوي صفرا لأننا في الواقع لم نر هذا التركيب في المدونة التي استنبطنا منها نحونا الثُّنائيّ. ولو لم نجد حلاًّ لهذه المشكلة فإنَّ هذا سوف يسبب ضررًا بالغا لأيّ استخدام لهذه النتائج، إذ إنَّ جملةً محتملة بصورة كبيرة، وربما بدرجة احتمال جملة (ذهب محمد إلى المدرسة)، لن نجدَ لها ما يدعمها من النَّحو الثُّنائيّ؛ والسَّبب أن المدوَّنات اللُّغَوِيَّة مهما كبرت فلن تغني عن أن واقع اللغات الحية متدفِّق ومتنامي. ويكادُ تعداد جمَلِ وتعبيرات هذا الواقع أن يكون لا نهائيًّا؛ فكيف نستنبط ما لم نره في المدونة؟

١ , ٤ – مشكلة: من خارج مُفردات المدَوَّنة (Out Of Vocabulary - OOV)

لو لم نجب على هذا السؤال ما أمكَنَ للنَّحو الإحصائيّ أن يكون مفيدا، لأنَّ ضرره سيكون أكبر من نفعه في كثير من الأحيان. وبعبارةٍ أخرى، لو لم يتمكَّن الباحثون من إيجاد حلول لهذه المشكلة لما كانت لهذا النحو قائمة.

تعالوا نفتَرض أن لدينا نحوًا فيه ١١ كلمة فقط، ووجدنا فيه ١١ حالة للنَّحو الثُّنائيّ يمكن أن نُقَدِّرَها تقريبا بـ ١١×١١=١٢١، أي أن هناك ١٢١ أيّ كلمة من مفردات

المدونة عقب كلمة أخرى (ويمكن أن تتكرر الكلمة، في مثل قوله تعالى: «وَجَاءَ رَبُّكَ وَالمَلَكُ صَفًّا صَفًّا»)، ولكن ورود ١٢ حالة فقط (كما فى الجدول رقم ٨- ٤) معناها أن هناك احتمالاً لـ ١٠٩ حالات لم ترد في المدونة. والحقيقة قد يكون ورود بعض التتابعات مستحيلا مثل ورود بداية جملة تتبعها بداية أو نهاية جملة .. إلخ، ولكن في مدونة حقيقية كبيرة لا يكون لهذه الاحتمالات أثر اذا اهملناها. وكذلك في الواقع الحقيقى يمكن أن نفرض أنه لن نرى إلا ~ ٥٠٪ من تتابع الكلمات بالنسبة لكل التتابعات الممكنة، هذا مقبول وحينئذ يمكن أن نحسب حساباتنا على توقع OOV ← ~ ٦٠ كلمة فقط. ولكن في مدونتنا البسيطة سنفرض للسهولة أن كل التتابعات ممكنة.

<u>حل المشكلة:</u>

لجأ كثير من الباحثين إلى محاولة تقدير احتمالات للمفردات والتَّتابُعات (الثنائية والثلاثية... إلخ) التي لم ترد في المدونة مع إعادة حساب التتابعات التي وردت بحيث يكون مجموع الاحتمالات واحدا صحيحا، لأن هذه من مسلمات نظرية الاحتمالات.

تعالوا نفترض أننا أضفنا مقدارًا ثابتًا، وقدره «٠,٠١» ، إلى كل احتمالات تتابع الكلمات؛ سوف نحتاج إلى إضافة ١٢١ مرة «٠,٠١» إلى البسط في ١٢١ حالة، شاهدنا فقط ١٢ حالة والباقي سنكتفي باعتبار وروده «٠,٠١» مرة تقديرا. ولذلك ستتغير الاحتمالات كما هو مبين في جدول رقم (٨-٤) العمود الأخير.

لنختبر نتائجَنا حتى الآن؛ هب أنَّنا سمعنا جملة، واختلط الأمر علينا بين جملتين:

- «**ذهب أحمد إلى المدرسة**».
- «**قابل إلى أحمد زميله**»

(لنرى معًا كيف يُستخدم النَّحو العدديّ لترجيح أقرب الحُلُول إلى الصَّواب).

بتطبيق نظرية الاحتمالات:

$$P\left(\text{ذهب أحمد إلى المدرسة}\right) \approx P_*\left(\text{ذهب}\right) * P_*\left(\text{ذهب/أحمد}\right) * P_*\left(\text{أحمد/إلى}\right) * P_*\left(\text{إلى/المدرسة}\right)$$
$$= \text{٠.٠٠٦٢٥} * \underbrace{\text{٠.٠٠٠٤٥}}_{(OOV)} * \underbrace{\text{٠.٠٠٠٤٥}}_{(OOV)} * \text{٠.٤٥٧} = \text{٥.٧٨} * \text{١٠}^{-\text{٧}}$$

بينما «**قابل إلى أحمد زميله**»

$$P\left(\text{قابل إلى أحمد زميله}\right) \approx P_*\left(\text{قابل}\right) * P_*\left(\text{قابل/إلى}\right) * P_*\left(\text{إلى/أحمد}\right) * P_*\left(\text{أحمد/زميله}\right)$$

$$= \text{٠.٠٠٦٢٥} * \underset{(OOV)}{\underline{\text{٠.٠٠٠٤٥}}} * \underset{(OOV)}{\underline{\text{٠.٠٠٠٣١}}} * \underset{(OOV)}{\underline{\text{٠.٠٠٠٤٥}}} = \text{٣.٩٢} * \text{١٠}^{-\text{٩}}$$

$$< P\left(\text{ذهب أحمد إلى المدرسة}\right)$$

إذن: تكون الجملة الأولى هي المَرَجَّحة.

٢- التنعيم (Smoothing)

تعالوا نعالج هذه المشكلة (من خارج المفردات) بطريقة أكثر عمقًا، تُعرَف بعملية التنعيم؛ أي: تنعيم قيم الاحتمالات الناتجة عن الحساب المباشر الناتج عن قسمة عدد التكرارات (سواء للكلمة أو الكلمتين المتجاورتين... إلخ) على العدد الكلِّيّ للكلمات في المدونة. وهناك طرق كثيرة للتنعيم نتعرف على أهمها.

٢, ١- التنعيم بالخصم (Smoothing by Discount)

كما أسلفنا فإن مشكلة عدم ورود كل الاحتمالات الممكنة في اللغة في قواعد البيانات المستخدمة في التدريب يسبب فشلا ذريعًا لاستخدام النحو العددي اذا لم تعالج هذه المشكلة. وهناك العديد من الطرق لتقدير هذه الاحتمالات.

- **تنعيم لابلاس (Laplace Smoothing)**

وتعتمد هذه الطريقة على تقدير عدد المرات التي نراها، ثم إضافة واحد لكل الحالات التي مرت بنا (بما في ذلك المرات التي مرت «صفر» مرة)؛ وبلغة الإحصاء:

$$P(w_j) = \frac{C_j}{N}$$

حيث C_j عدد مرات ورود الكلمة w_j، و N العدد الكلي لكلمات المدونة.

وتصبح بعد طريقة تنعيم لابلاس:

$$P_{Laplace}(w_j) = \frac{C_j + 1}{N + V}$$

حيث V عدد المفردات المختلفة التي يمكن أن نصادفها. ولنأخذ مثالا لذلك:

هب أننا نملك مدونة بها ٣ كلمات فقط، وقدَّرنا أن هناك كلمة واحدة يمكن إضافتها؛ إذن ستكون (V=٤). وبافتراض ورود الكلمات كالآتي:

عدد ورود الكلمة بعد التنعيم	عدد ورود الكلمة قبل التنعيم	
٤	٣	$C_1=C(w_1)$
٣	٢	$C_2=C(w_2)$
٢	١	$C_3=C(w_3)$
١	٠	$C_4=C(w_4)$
$V_2= 4, N_2=10$	$V_1= 3, N_1=6$	

قبل التنعيم: N_1 في هذه الحالة = ٦، و V_2 = ٣ (مفردات):

$$P(w_1) = \frac{C_1}{N_1} = \frac{3}{6} = 0.5$$

$$P(w_2) = \frac{C_2}{N_1} = \frac{2}{6} = 0.33$$

$$P(w_3) = \frac{C_3}{N_1} = \frac{1}{6} = 0.167$$

لتصبح بعد تنعيم لابلاس $N_1 \leftarrow N_2$ = ١٠، و $V_1 \leftarrow V_2$= ٤ (مفردات):

$$P_{Laplace}(w_1) = \frac{3+1}{6+4} = 0.4$$

$$P_{Laplace}(w_2) = \frac{2+1}{6+4} = 0.3$$

$$P_{Laplace}(w_3) = \frac{1+1}{6+4} = 0.2$$

$$P_{Laplace}(w_4) = \frac{0+1}{6+4} = 0.1$$

وإذا جمعت كل الاحتمالات الآن سوف تجد أنها تساوي الواحد الصحيح، بما يتفق مع إحدى مسلمات نظرية الاحتمالات.

من الواضح أن إضافة واحد صحيح لكل مرات ورود المفردات يضعف بشكل ملموس احتمال المفردات التي وردت في المدونة بتكرار قليل بالنسبة لتلك المفردات التي لم ترد على الإطلاق؛ لذلك فإن هناك محاولات لتحسين هذا النوع من التنعيم بإضافة كمية ثابتة أقل من الواحد، وهذا يعتمد على حجم المدونة المستخدمة للتدريب.

ولكن كيف يتم تقدير عدد المفردات V؟ بالنسبة للنحو الأحادي، يتم تقديره على أساس المعرفة باللغة؛ ولكن اللغة العربية غنية جدا في عدد كلماتها؛ ففي مدونة من حوالى ١٥٨ مليون كلمة من الأخبار وجدنا بها ٩٥٠ ألف مفردة مختلفة بعضها عن بعض (كتاب، والكتاب، تُحسبان كلمتين مختلفتين)، وفي مدونة قريبة من ٦٦٠ مليون كلمة وجدنا قريبًا من ١,٨ مليون مفردة، لكنَّها احتوت على كمية كبيرة من الأخطاء اللغوية. فنحن نُقَدِّر المفردات الصحيحة في هذه الحالة بنحو ١,٤ مليون كلمة. لذلك عند التعامل مع مجال مثل الأخبار (وبالمناسبة، هو من المجالات الغنية بالمفردات لكثرة مجالاته الفرعية من سياسة واقتصاد، ورياضة، وعلوم، وحالات الطقس... إلخ) يمكن فرض أن عدد المفردات التي نتعامل معها قد يصل إلى أكثر من ٢ مليون مفردة، مع مُلاحظة أن اسم قرية جديدة أو مدينة وقَعَ بها زلزالٌ يضيف مفردة جديدة للمجال كل يوم.

ملاحظة: ليس بالضرورة أن تكون إضافة ١ هو الحل الوحيد المتاح إذ يمكن إضافة كمية ثابتة أقل - كما في المثال الذي سقناه آنفًا (وإن لم يكُن بالضرورة منخفضًا جدًّا كما فعلنا، إنما اخترنا القيمة القليلة (٠,٠١) لتناسب بساطة المدونة المستخدمة). وعادة ما يتم ذلك عبر عدة تجارب.

ومن الجدير بالذِّكر أننا في مثل هذه التجارب نحتاج إلى تقسيم المدونة إلى ٣ أقسام:

- القسم الأول للتعلم (في حالتنا لتعلم النحو العدديّ).
- القسم الثاني لاختيار أفضل القيم لبعض المعاملات (في حالتنا لاختيار أفضل قيمة للثابت (١، ٠,٥، ٠,١...).
- القسم الثالث للاختبار النهائيّ، ولا يجوز تغيير المعاملات ثم إعادة التجربة، لأن ذلك يعني أننا استعملنا قسم الاختبار في التدريب. لتوضيح ذلك، هب أننا أعددنا اختبارا

للطلاب فوجدنا مستواهم ضعيفًا في موضوع ما، فراجعناه معهم ثم أعدنا لهم نفس الامتحان! هذا لا يفرز الطالب الحافظ من الطالب الفاهم، لهذا الغرض خُصِّصَ القسم الثاني لاغراض ضبط متغيرات الحل.

- **خصم جود-تيورينج (Good-Turing Discount)**

وهي نظريَّةٌ إحصائيَّة، تُنسَب إلى العالمين «إرفنج جود (Irving John Good) وألان تيورينج (Alan Turing)» . وتعتمد منهجية الخصم هنا على فكرة بسيطة. إذا حسبنا عدد المفردات التي وردَت في المدونة مرة واحدة، ولنسمها N_1 ، وعدد المفردات التي وردَت في المدونة مرتَين، ولنسمها N_2، وهكذا سنحصل على N_3 ، N_4 ، N_5،....

وكذلك يمكن تقدير N_2 أي.. المفردات التي لم ترد في المدونة – ولو تقديرًا نظريًّا؛ فإننا لو افترضنا في تخصص معين أننا لن نتجاوز المليون مفردة، فإن

$$N_0=1,000,000 - N_1 - N_2 - N_3 - N_4 \ldots\ldots$$

ونعود لمنهجية تقدير احتمالات ورود المفردات:

$$C^* = (C+1)\frac{N_{C+1}}{N_C}$$

حيث C هو عدد التكرارات الحقيقي، و *C هو التكرار التقديري لأغراض تنعيم الاحتمالات. ومن الملاحَظ في أيَّة مدونة أنه كلما زادت تكرارات المفردات كُلَّما قَلَّت أعدادها؛ وهذا يعني أن $N_c > N_{c+1}$ (هذه العلامة > تعني أن شمالها أكبر من يمينها). ولذلك يمكننا اعتبار أن $\frac{N_{C+1}}{N_C}$ هي مقدار التخفيض في الأعداد. ولو لاحظت أننا زدنا «١» وخفضنا بمقدار $\frac{N_{C+1}}{N_C}$ فستكون النتيجة:

- تخفيض في قيم الاحتمالات لما ورد من مفردات المدونة.
- وجود قيمة لاحتمالات ورود المفردات التي لم ترد في المدونة.

تعالَ نستدعي مدونتنا الصغيرة مرة أخرى:

- **ذهب محمد إلى المدرسة.**

- **حين وصل محمد إلى المدرسة قابل زميله أحمد.**

في مدونتنا السابقة؛ كما ورد منها في مدونتنا البسيطة ١١ كلمة.

إذن: تكون الأعداد N_c كالآتي:

N_0 (عدد المفردات التي لم ترد في المدونة) = ١٢١ – ١١ = ١٠٩

N_1 (عدد المفردات التي وردت مرة واحدة) = ٦

N_2 (عدد المفردات التي وردت مرتين) = ٥

محمد، إلى، المدرسة، بداية الجملة ونهاية الجملة، ولا تنس أن عدد الكلمات الكلِّي المشاهد في المدونة هو ١٦ كلمة N = . وعليه، سيكون تطبيق منهجية التنعيم باستخدام جود–تيورينج في تقدير احتمال النحو الثنائى الذى لم نره في المدونة:

$$P_{GT}\ (\text{لأي تتابع لم نره}) = \frac{(0+1)\frac{N_1}{N_0}}{N} = \frac{6/109}{16} = 0.00344$$

والرمز $P_{GT}(x)$ يعني احتمال ورود (x) بتنعيم جود–تيورينج.

٢,٢– التنعيم باستخدام الإدراج (Interpolation)

ترتكز طرق التنعيم بالخصم على تقدير قدر مناسب من الاحتمالات للحالات التي لم نر فيها خصمًا مما ورد علينا في المدونة. ولكن التنعيم بالإدراج يفيد في حسن تقدير ما ورد علينا في المدونة، وذلك كالآتي؛ إذا أردنا تحسينا للنحو الثلاثي مثلا:

$$\hat{P}(w_n / w_{n-1}\, w_{n-2}) = \lambda_1 P(w_n / w_{n-1}\, w_{n-2}) + \lambda_2 P(w_n / w_{n-1}) + \lambda_3 P(w_n)$$

بحيث يكون

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

وقيم λ_j أعلاهُ يمكن إيجادها بإجراء التجارب ووضع قيم مختلفة لها واختيار القيم التي تعطي أفضل النتائج للنحو (حسب المشكلة المستخدم فيها النحو).

ولقراءة هذه المعادلة لتكون مفهومة أكثر سنعيد كتابتها بالكلام:

الاحتمال المقدر للكلمة ن بشرط ورود الكلمتان ن-١، ن-٢=

ثابت يقدر من مدونة التدريب * احتمال كلمة ن بشرط ورود الكلمتان ن-١، ن-٢ قبلها

ثابت آخر يقدر من مدونة التدريب * احتمال كلمة ن بشرط ورود كلمة ن-١ قبلها

ثابت آخر يقدر من مدونة التدريب * احتمال كلمة ن (أى النحو العددى)

على أن مجموع الثوابت الثلاثة لابد أن يكون واحدا صحيحا.

والفكرة من وراء هذا التحسين لتقدير النحو العددي تتبين من هذا المثال:

«**قال الله تعالى**» و «**رضي الله عنه**».

نفترض جدلاً أننا عند دراسة النحو الثلاثي لكلمتي (تعالى)، (وعنه) وجدنا أن تكرارهما متساوٍ في المدونة؛ ولكن كان ورود (الله تعالى) أكثر من (الله عنه)؛ وعليه.. فسيُساهم هذا في رفع احتمال (قال الله تعالى) عن (رضي الله عنه).

٢,٣- التنعيم بالتراجع (Smoothing using back-off)

- **تراجع كاتز (Katz back-off)**

يُستَخدَم تراجُع كاتز - عادةً - كمُكمِّلٍ لخصم جود-تيورينج؛ وتُستَوحى فكرتُه من التنعيم بالإدراج؛ ويمكن من خلاله فهم كيفية تقدير النحو العددي من درجة أعلى بدلالة النحو العددي من الدرجة الأدنى منه مباشرةً في المدونة.

لو أن عندنا نحوا ثلاثيا مطلوب تقديره، لأننا لم نره في المدونة، فإنَّ الطرق السابقة للخصم - وربما أفضلها حتى الآن جود-تيورينج -ستعطي كُلَّ ما لم نره نفسَ الاحتمال، ولكن «كاتز» يُقدِّرها اعتمادا على النحو الثنائي والأحادي إذا لزم الأمر. وهذا يعني أن

نعطي احتمالا أكبر للنحو الثلاثي المقدر لكلمات لم ترد في المدونة إذا كان نحوها الثنائي أكبر. ولبسط التعريف الرياضى انظر اسفل الصفحة[١].

- **التنعيم باستخدام طريقة «نزر–ناي» (Kenser-Ney)**

نستطيعُ الوُقُوفَ على هذه الطَّريقة من خلال المثال التَّالي:

أردت أن **أقرأ فأخرجت**...... ولم يرد في المدونة مثل هذه الجملة قط

بافتراض وجود كلمتين مرجحتين ولهما نفس النحو العددي الأقل، هما:

- «**النظارة**» (ما ورد في المدونة: عملت النظارة، وقعت النظارة، استخدمت النظارة، وضعت النظارة،....).
- «**بور**» والتي لم ترد إلا في (بور سعيد، بور فؤاد).

فإن كلمة «النظارة» تُرجَّح، لأن ورودها مع كلمات أكثر في المدونة يجعلها مرشحة للورود أكثر من كلمة «بور» فيما لم نره[٢].

١- نحتاج أن نعّرف:

أي تكرارات "x" C(x)= count of x

احتمال x بعد الخصم (باستخدام طريقة من طرق الخصم السابقة) P*(x)

وبدلا من استخدام w_n, w_{n-1} , w_{n-2} فإننا سنستخدم x,y,z لتكون المعادلات كالآتي:

$$P_{katz}(z/x,y) = \begin{cases} P^*(z/x,y) & if\ C(x,y,z) > 0 \\ \propto(x,y)P_{katz}(z/y), & else\ C(x,y) > 0 \\ P^*(z), & otherwise \end{cases}$$

$$P_{katz}(z/y) = \begin{cases} P^*(z/y), & if\ C(y,z) > 0 \\ \propto(y)P^*(z), & otherwise \end{cases}$$

حيث ∝ تعني معامل التطبيع (لتجعل مجموع الاحتمالات ١ صحيحًا)، ولنقل اعتماد النَّحو العدديّ من درجة أعلى إلى درجة أقل. أما (x,y) أو (y) ∝ فتعني أن هذا المعامل متغير يعتمد على ما بين الأقواس.

وتجدُرُ الإشارةُ إلى أنَّ تراجع كاتز يمكن تعميمه على أيّ درجة من النحو العددي؛ أي أنَّ اقتصار المعادلات التي ذكرناها على النحو العددي من الدرجة الثالثة هو لمجرد التبسيط وتوضيح الفكرة. كذلك فإنَّ المعاملات ∝ يجري حسابها أيضًا من تكرارات النحو الأحادي والنَّحو الثنائي ... إلخ.

١- وتصاغ معادلاته كالآتي:

$$P(w_i/w_{i-1}) = \frac{C(w_{i-1}w_i)-d}{C(w_{i-1})} + \beta(w_i)\frac{|\{w_{i-1}:C(w_{i-1}w_i)>0\}|}{\sum_{w_i}|\{w_{i-1}:C(w_{i-1}w_i)>0\}|}$$

حيث d ثابت يطرح من كل احتمال لنحو ثنائيّ ورد في المدونة.

و $\beta(w_i)$ تُختار (وهي مختلفة من كلمة لأخرى) لتجعل مجموع الاحتمالات ١ صحيحًا.

و $|\{w_{i-1}:C(w_{i-1}w_i) > 0\}|$ تعني عدد الكلمات المختلفة (w_{i-1}) التي ترد فيها مع في المدونة، مع ملاحظة أننا نحصي التنوع وليس عدد مرات الورود. مثال: لو وردت الكلمة (w_i) ١٠ مرات مع كلمة و٥ مرات مع كلمة أخرى فقط، فيكون مفهوم التعبير الرياضيّ المذكور هو ٢ وليس ١٥. (حيث يشير التعبير الرياضيّ |....| إلى أنَّ القيمة المذكورة تشمل عدد الأنواع، وليس عدد التَّكرارات).

و $\sum_{w_i}|\{w_{i-1}:C(w_{i-1}w_i) > 0\}|$ تعني مجموع عدد المرات التي وردت فيها كلمات مختلفة في المدونة كلها.

٣- موضوعات تساعد على تحسين النَّحو العدديّ

هناك بعض الجوانب الَّتي تساعد على تحسين التقدير، ومن ذلك:

٣,١ - النحو العددي الفئوي (Class Based N-gram)

خذ هذه الأمثلة:

كان راتب سعيد ١٠٠٠ جنيه في الشهر

ذهب على إلى الإسكندرية يوم الأربعاء

ركبت مريم طائرة مصر للطيران

فلو ارتبط النحو العددي برقم (١٠٠٠) فقط لما استفدنا من هذه المعلومة لو جاء الرَّاتب مختلفا في موضع جديد؛ ولكن يمكن أن نحدد أن هناك فئة من الأرقام يمكن أن يحل أحدها مكان الآخر. وكذلك أيام الأسبوع أو الشهور أو أسماء شركات الطيران... إلخ.

ففي المدونات قليلة العدد يمكن تعظيم الفائدة منها إذا عالجنا بعض الأسماء والأرقام باستخدام اسم الفئة التي تنتمي إليها هذه الأسماء أو الأرقام.

٣,٢ - النحو العدديّ الموضوعيّ (Topic Based N-gram)

تتأثر النتائج كثيرا بشكل إيجابي إذا استخدمنا نحوًا عدديًّا من مدونة ذات موضوعات مشابهة للموضوع الذي نحن بصدده.

لذلك يمكن حساب النحو العددي لمدونات تحتوي كلٌّ منها على موضوعات متشابهة، مثل (مدونة سياسية، اقتصادية، علمية، قانونية،... إلخ). وهناك إضافات نوعية قد تكون مفيدة عند استخدام النحو العددي، ومنها الاستفادة من ظاهرة: الاستدعاء.

٣,٣- دعم النَّحو العدديّ بالاستفادة من ظاهرة الاستدعاء

خذ هذا المثال:

ذهب إلى

ذهب محمد إلى

ذهب محمد وعلي إلى

ذهب محمد وعلي وسمير إلى

تلاحظ أن كلمة «ذهب» استدعت وجود كلمة «إلى» في كثير من الأحيان بعدها.

٣,٤ - النَّحو العدديّ متغير الطول (Variable length N-gram)

للنَّحو العدديّ أهمية قصوى في تطبيقات كثيرة؛ ولذلك نحتاج إلى دعمه بنظريات جديدة لغوية المنشأ مستوعبة لاحتياج الحاسوبيين، وخاصة مع اللغة العربية الَّتي تتمتَّعُ بظاهِرَتَي الاشتقاق والتَّوليد. وإذا كنَّا نحتاج في كثير من التطبيقات، مثل: التعرف على الكلام المنطوق في اللغة الإنجليزية، إلى ٦٤ ألف كلمة تغطي ٩٩٪ من احتياجات الكلمات في مجال معين (مثل مجال الأعمال Business) فإننا نحتاج إلى أكثر من ٦٠٠ ألف كلمة عربية لنقرب من درجة التغطية ٩٩٪. إن ذلك يجعل احتياجنا لمدونات كبيرة جدًّا لا مفر منه، والاحتياج إلى المعالجات اللغوية المسبقة ضرورة. ومن هذه المعالجات التحليل الصرفي لمعرفة السوابق واللواحق وجذع الكلمة، وربما نحتاج أيضا للوزن والجذر. (اللافت للانتباه أن العربية مبنية بعدد محدود من السوابق واللواحق والأوزان والجذور) إلا أن بناء النحو من هذه اللبنات له تحدياته ويستغرق جهودًا علميَّةً عميقة من اللغويين والحاسوبيين للخروج بنحو عددي يستفيد من ميزات اللغة العربية وتطورها الصرفيّ، ويلبي حاجة التطبيقات المختلفة.

٤ - تقويم قوة النَّحو العَدَدِيّ

نحتاج إلى تقويم كفاءة النحو المستخدم، ففي بعض التطبيقات يقيسون هذه الكفاءة بما يسمى مقدار «الالتباس» (Perplexity). وكلما قل الالتباس يعني ذلك كفاءة أعلى للنحو المستخدم. ويحسب مقدار الالتباس كما في المثال التالي:

على سبيل المثال، في اللغة الإنجليزية يحسب الالتباس عندما لا يكون هناك نحو على الإطلاق في تقنية التَّعَرُّف على الكلام المنطوق لعدد كلماتٍ مُحتَمَلة تدرَّبَت عليها التقنية، ومقدارها ٢٠٠٠٠ كلمة.

فكان مقدار الالتباس كما هو مُبَيَّنٌ في الجدول الآتي:

النحو العدديّ (N-gram)	الالتباس (Perplexity)
بدون نحو على الاطلاق	٢٠٠٠٠
النحو الأحادي (Uni-gram)	٩٦٢

النحو العدديّ (N-gram)	الالتباس (Perplexity)
النحو الثنائي (Bi-gram)	١٧٠
النحو الثلاثي (Tri-gram)	١٠٩

لننظر كيف انخفض مقدار الالتباس من ٢٠٠٠٠ بدون أي معلومات معطاة للنظام عن اللغة، إلى فقط ١٠٩ بعد استخدام النَّحو الثُّلاثِيّ. يمكن النَّظر إلى هذه الأرقام كالآتي: كأنّ المهمة التي تلقى على عاتق النظام قبل إعطاءه أي معلومات لغوية عند التعرف على الكلمة التي سمعها هي مهمة اختيار كلمة من ٢٠٠٠٠ كلمة. وليس له دليل على هذه الكلمة إلا ما يسمعه من صوت. وتنخفض درجة الالتباس لنفس المهمة إذا أفدنا النظام بمعلومات عن اللغة واستخداماتها وتتابعات كلماتها ملخصة في النحو الثلاثي لتصبح المهمة كما لو كانت هي التعرف على كلمة من ١٠٩ كلمة فقط باستخدام المعلومات الواردة من الصوت. هل نستطيعُ تصوُّرَ النتائج في الحالتين؟ الحالة الأولى: يفشل النظام تماما في الوصول إلى نتيجة لها أي اعتبار، أما في الحالة الثانية فإن النتائج يمكن أن تزيد عن ٩٠٪ كنسبة دقَّة في التعرف على الكلام المنطوق في ظروف مناسبة.

فبالرغم من بساطة فكرة النحو العدديّ إلا أنه – وبعد المعالجات المختلفة لما لم يره من كلمات وتتابعات – أصبحَ مُفيدًا للغاية وعمليا إلى درجةٍ كبيرة.

هل لديك أخي الباحث فكرة نيرة كهذه يصلح مع تطبيقها أن نصل لنتائج أفضل؟ إذا أمكن تمثيلُ اللُّغة رياضيًّا، فإنَّنا كعاملين في مجال تقنيات اللُّغة سنستفيد كثيرا من ذلك. **فشمِّر واجتهِد.**

وهناك العديد من الأعمال الآن في مجال توليد نماذج لغوية من الشبكات العصبية؛ والنتائج تتحدث عن تفوق ملموس عن النحو العددي، إلا أنها تحتاج لحسابات تأخذ في الغالب وقتا أطول بكثير من ذلك الوقت الذي يحتاجه النحو العددي.

٥- أمثلة على مجالات الإفادة من النَّحو العَدَدِيّ

١- **التَّعرُّف على الكلام المنطوق**؛ كما أسلفنا. فربما كان هذا هو التطبيق الأول الذي أظهر قوة النَّحو العدديّ وتمَّ من خلاله علاج أخطر مُشكلاته، وهي عدم رؤيته لحالات كثيرة محتمَلة.

٢- **التَّدقيق الإملائيّ**؛ ولعلَّنا نُلاحظُ إشارات الخطأ الحمراء الَّتي يُنَبِّهُنا إليها البرنامج المكتبيّ «ميكروسوفت ورد MS-Word» ، وما يُرفقه من احتمالاتٍ للصَّواب. إن أصل العمليَّات التي يقوم بها هذا **المدقق الإملائي** هي من مثل النَّحو العدديّ.

٣- **الترجمة الآلية**؛ فقد تطورت نُظُم **الترجمة الآلية**، وأمكنَ من خلالها توليد عباراتٍ أكثر دقَّة عند استخدام النَّحو العدديّ في توليد الترجمة للُّغــة المستهــدفة.

٤- كما أن هناك في ساحة **محركات البحث** فرصة لتحسين البحث باستخدام النحو العدديّ.

٥- وكذلك في التطبيقات التعليمية **لتعليم اللغات** حيث يُستخدم الحاسب لتحليل ما كتبه المتعلم والحكم عليه. وهنا أيضا يستفاد من النحو العدديّ.

٦- هناك **نظم للتعرف على الحروف العربية**، فمنها المصمم للتعرف على الكلام المطبوع، ومنها المصمم للتعرف على الكلام المكتوب باليد، ولولا استخدامُ النحو العدديّ في هذه التَّطبيقات لكانت النتائج جد هزيلة ...

٦- أفكارٌ بحثيَّة لأطرُوحاتٍ علميَّةٍ مُستَقبليَّة

١- تكوين مدونة لبعض المجالات، تُختار موضوعاتها بحيث تحقق أعلى تغطية للكلمات التي يمكن أن تأتي في هذا المجال.

٢- البحث في أفكار جديدة لمعالجة مشكلة الكلمات التي لم نرها من قبل (في المدونة المخصصة للتدريب)، والتي نسميها التنعيم. كلما استفدنا من خصائص اللغة كلما كانت الحلول أوفق وأفضل.

٣- تكوين موارِد لغوية تساعد على تفضيل كلمة عن كلمة أخرى متقاربتين في النطق أو الكتابة اعتمادًا على خصائص دلالية للكلمتين.

٤- عمل معاجم مستنبطة من مدونات ترجح استخدام كلمة عن كلمة متقاربة معها فى الرسم أو النطق تبعا للسياق.

٥- تحتاج كثير من التطبيقات كالتعرف على الكلام المنطوق إلى معرفة نطق الكلمة الصحيح من سياقها - فوضع منظومة من القواعد المساعدة لضبط الكلمة من سياقها سيساعد كثيرا.

ببليوجرافيا مرجعيَّة

1. Bellegarda, J. R. & Monz, C. (2016). State of the art in statistical methods for language and speech processing. Computer Speech & Language, Elsevier, Vol. 35, pp. 163-184.
2. Cui, J. (2011). Integrating Linguistic and Statistical Knowledge in Language Modeling. BiblioBazaar.
3. Deng, L.; Liu, Y. (2018). Deep Learning in Natural Language Processing. Springer.
4. Farghaly, A. A. S. (2010). Arabic Computational Linguistics. University of Chicago Press.
5. Franz, A. & Brants, T. (2006). "All Our N-gram are Belong to You". Google Research Blog.
6. Friedenthal, S. & Moore, A. & Steiner, R. (2011). A Practical Guide to SysML: The Systems Modeling Language. Elsevier.
7. Goutte, C. (2009): Learning Machine Translation. MIT Press.
8. Huang, G. & Huang, G.B. & Song, S. & You, K. (2015). Trends in extreme learning machines: a review. Neural Networks, Elsevier, Vol. 61, pp. 32-48.

9. Johnson, M. & Khudanpur, S. P. & Ostendorf, M. & Rosenfeldm R. (2004). Mathematical Foundations of Speech and Language Processing. Springer.

10. Jozefowicz, R. & Vinyals, O. & Schuster, M. & Shazeer, N. & Wu & Y. (2016). Exploring the limits of language modeling. Cornell University.

11. Jurafsky, D. & Martin, J. H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall.

12. Koren, B. & Vuik, K. (2010). Advanced Computational Methods in Sci-ence and Engineering. Springer.

13. Kumarm E. (2011). Natural Language Processing. I. K. International Pvt Ltd.

14. Luong, M. T.& Le, Q. V. & Sutskever, I. & Vinyals, O. & Kaiser, L. (2015). Multi-task sequence to sequence learning. Cornell University.

15. Lv, Y. & Zhai, C. (2009). Positional Language Models for Information Retrieval, in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR).

16. Manning, C. D. & Schütze, H. (1999). Foundations of Statistical Natural Language Processing, MIT Press: ISBN 0-262-13360-1.

17. Matsumoto, Y. & Sproat, R. & Wong, K. & Zhang, M. (2006). Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead: 21st International Conference, ICCPOL 2006, Singapore, December 17-19, 2006, Proceedings.

18. Mishra, S. (2018). Artificial Intelligence and Natural Language Processing. Cambridge Scholars Publisher.

19. Mulder, W. D. & Bethard, S. & Moens, M. F. (2015). A survey on the application of recurrent neural networks to statistical language modeling. Computer Speech & Language, Elsevier, Vol. 30, pp. 61-98.

20. Neamat El, G.; Yee, S. (2018). Computational Linguistics, Speech and Image Processing for Arabic Language. World Scientific.

21. Olive, J. (2011). Handbook of Natural Language Processing and Machine Translation. Springer.

22. Sauro, J. & Lewis, J. R. (2012). Quantifying the User Experience: Practical Statistics for User Research. Elsevier.

23. Schimek, M. G. (2012). Smoothing and Regression: Approaches, Computation, and Application. John Wiley & Sons.

24. Soudi, A. (2012). Challenges for Arabic Machine Translation. John Benjamins Publishing.

25. Srinivasa-Desikan, V. (2018). Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing.

26. Su, Y. (2011). Knowledge Integration into Language Models: A Random Forest Approach. BiblioBazaar.

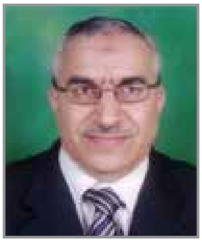
27. Sundermeyer, M. & Ney, H. & R Schlüter, R. (2015). From feed-forward to recurrent LSTM neural networks for language modeling. IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 23, pp. 517-529.

28. Wei, X. (2007). Topic Models in Information Retrieval. ProQuest.

29. Weinert, H. L. (2013). Fast Compact Algorithms and Software for Spline Smoothing. Howard L. Weinert.

30. Zhai, C. (2009): Statistical Language Models for Information Retrieval. Morgan & Claypool Publishers.

الباحثون



الدكتور/ مُحسِن عبد الرَّازق علي رشوان

يشغل منصب أستاذٍ بقسم الإلكترونيات والاتصالات الكهربائية في كُلِّيَّة الهندسة - جامعة القاهرة. تخرَّجَ عام ١٩٧٧ وكان الأول على دفعته، وحصلَ على ثلاثة ماجستيرات، ثم على الدُّكتوراه من جامعة كوين بكندا؛ أشرف على أكثر من مائة رسالة ماجستير ودكتوراه. يدير الشَّركة الهندسيَّة لتطوير النُّظُم الرَّقمِيَّة RDI المتخَصِّصة في مجال تقنيات اللُّغة العربيَّة.



الدكتور/ المُعتزّ بالله السَّعيد طه

أستاذ الدِّراسات اللُّغويَّة الـمُساعد بجامعة القاهرة، وأستاذ اللِّسانيَّات الحاسوبيَّة المـُشارك بمعهد الدَّوحة للدِّراسات العُليا، ومُنسِّق وَحدة الموارد الـمُعجميَّة بمشروع مُعجم الدَّوحة. نَشَرَ نحوَ ثلاثينَ ورقة علميَّةٍ، بالإضافة إلى عددٍ من الكتب في الـمُعجَمِيَّة العَرَبيَّة والدِّراسات اللُّغَوِيَّة الـمُعاصِرة، وأسهمَ في أكثر من عشرة مشرُوعاتٍ بحثيَّةٍ دوليَّةٍ في ميادين مُعالَجَة اللُّغات الطَّبيعيَّة. حصلَ على عددٍ من الجوائز في مَيدان تخصُّصه، منها: جائزة (ألكسو ALECSO) للإبداع والابتكار في «المـَعلُوماتيَّة والـمُعالَجَة الآليَّة للُّغة العربيَّة»، وجائزة راشد بن حُمَيد للعُلُوم والثَّقافة.

الدكتور/ عبد العاطي إبراهيم هوَّاري



عمل باحثًا زائرا في جامعة جورج واشنطن، في الولايات المُتَّحدة الأمريكيَّة. حصل على درجة الدُّكتوراه في اللسانيات عام ٢٠٠٨م. عملَ في العديد من المشروعات البحثيَّة العربيَّة؛ كما عملَ باحثًا في جامعة كولورادو وجامعة كولومبيا الأمريكيَّة قبلَ أن يتَّجِّهَ للعمل في جامعة جورج واشنطن. نَشَرَ عددًا من الأوراق البحثيَّة في الدِّلالة المُعجميَّة وقضايا المُعجميَّة العَرَبِّيَّة والصرف العربي، كما شاركَ في العديد من المؤتمرات الدَّوليَّة داخلَ مصر وخارجها. له عددَ من المُؤَلَّفات العلميَّة المنشورة.

الدكتور/ سامح سَعد أبو المَجد الأنصاريّ



يَعمل أستاذًا للِّسانيَّات الحاسُوبيَّة ورئيسًا لقسم الصَّوتيَّات واللِّسانيَّات بكُلِّيَّة الآداب بجامعة الإسكندريَّة، ومديرًا لمركز اللُّغويَّات الحاسوبيَّة العربيَّة بمكتبة الإسكَندَرِيَّة. شارك في العديد من المشروعات العلمية ونَشَرَ العَدِيدَ من الأوراق البَحثِيَّة المَعنِيَّة بحوسبة اللُّغة العَرَبِيَّة؛ وهو عُضوٌ بجمعيَّة اللِّسانيَّات العَرَبيَّة بالولايات المُتَّحدة الأمريكِيَّة، وعضوٌ بمُؤَسَّسة لُغة الشَّبكات الدلالية الحاسوبية العالمِيَّة بجنيف.

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

مباحث لغوية ٥١

الموارد اللُّغويَّة الحاسوبيَّة

يُصدِر مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية هذا الكتاب ضمن سلسلة (مباحث لغوية)، وذلك وفق خطة عمل مقسمة إلى مراحل، لموضوعات علمية رأى المركز حاجة المكتبة اللغوية العربية إليها، أو إلى بدء النشاط البحثي فيها، واجتهد في استكتاب نخبة من المحررين والمؤلفين للنهوض بعنوانات هذه السلسلة على أكمل وجه.

ويهدف المركز من وراء ذلك إلى تنشيط العمل في المجالات التي تُنَبّه إليها هذه السلسلة، سواء أكان العمل علميا بحثيا، أم عمليا تنفيذيا، ويدعو المركز الباحثين كافة من أنحاء العالم إلى المساهمة في هذه السلسلة.

وتودّ الأمانة العامة أن تشيد بجهد السادة المؤلفين، وجُهد مُحرِّرَي الكتاب، على ما تفضلوا به من رؤى وأفكار لخدمة العربية في هذا السياق البحثي.

والشكر والتقدير الوافر لمعالي وزير التعليم المشرف العام على المركز، الذي يحث على كل ما من شأنه تثبيت الهوية اللغوية العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة.

والدعوة موجّهة إلى جميع المختصين والمهتمين للتواصل مع المركز؛ لبناء المشروعات العلمية، وتكثيف الجهود، والتكامل نحو تمكين لغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة.

الأمين العام للمركز

أ.د. محمود إسماعيل صالح

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for The Arabic Language

ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

9 786038 221549